A Framework for Quality Evaluation of VGI linear datasets

Thomas Koukoletsos

Thesis submitted for the Degree of

Doctor of Philosophy (PhD)

University College London (UCL)

March, 2012

Author's Declaration

I have read and understood the College and Department's statements and guidelines concerning plagiarism.

I declare that all material described in this thesis is all my own work except where explicitly and individually indicated in the text. This includes ideas described in the text, figures and computer programs.

All maps / data from Ordnance Survey within Digimap are used according to the following copyright:
© Crown Copyright / database right 2012. An Ordnance Survey/EDINA supplied service.

All maps / data from OpenStreetMap are used according to the following copyright:
Map data © OpenStreetMap contributors (www.openstreetmap.org), CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.0/)

Abstract

Spatial data collection, processing, distribution and understanding have traditionally been handled by professionals. However, as technology advances, more spatial data are needed for an increasing number of applications, sometimes of new and non-traditional types. Higher educational levels and interaction with new technologies equip the public with a better geographic understanding. By using technological and web developments, non-experts can now collect Geographic Information (GI), create spatial databases and distribute GI through web applications. With the help of volunteers, these datasets are dynamically updated and are provided at lower or no cost, compared to the official spatial sources. Professionals, on the other hand, are striving to update their datasets in an environment of increasing competitiveness and reduced funding, and these new crowd-sourced data options that seem to be a threat, can also be an opportunity, if appropriately handled.

There are various areas that could benefit from this Volunteered Geographic Information (VGI). Local authorities and other organisations could use VGI because it provides more localised and customised data. Professionals could use it to update their datasets or create new ones to satisfy the contemporary spatial requirements. Crisis management has already proved to benefit from VGI, as the effort of hundreds of volunteers cannot be matched by the work of the official sources' limited personnel to update GI in an area hit by a natural disaster. Developing countries have no or unreliable official datasets and VGI could aid in the provision of the basic infrastructure services. Non-governmental organisations or individuals that cannot afford the cost of official datasets could also use VGI for their spatial needs.

The most concerning issue with VGI is its unknown quality. Usually there are no standards or metadata, unknown methods of collection, data density that depends on the number and dedication of the users within an area and none can be held responsible for the data provided. These factors lead to heterogeneous datasets that traditional quality measurement methods cannot handle. In any case, the quality elements that these methods measure were standardised long before the appearance of VGI, so not all of them are applicable. The frequency of updates also renders any quality results obsolete shortly after the analysis. The lack of a quality framework with an appropriate level of automation, which would enable the repetition of the VGI quality assessment, renders the choice of using it difficult or risky for potential users.

This thesis proposes a framework for quality evaluation of linear VGI datasets, used to represent road or other networks. Linear data are the majority of spatial data in a map, but their nature does not allow for a standardised evaluation method, despite numerous approaches already proposed by researchers.

The suggested automated methodology is based on a comparison of a VGI dataset with a reference one (dataset of known quality). The heterogeneity issue is handled by producing individual results for small areal units, using a tessellation grid. The quality elements that are considered important for VGI are then measured. These include data completeness, attribute and positional accuracy.

Compared to previous research in VGI, this thesis includes an automated data matching procedure which is specifically designed for VGI. It combines geometric and thematic constraints, shifting the scale of importance from geometry to non-spatial attributes, depending on their existence in the VGI dataset. Based on the data matching results, all quality elements are then measured for corresponding objects, which provides a more accurate quality assessment.

Data completeness and attribute accuracy calculation is based on the length of corresponding features, because length proves to be more useful as a quality indicator. Positional accuracy uses an already proposed buffering method (Goodchild and Hunter, 1997), however this is the first time it is applied in its suggested form. All quality results refer to the tile level, however quality information is also stored at feature level for data completeness and attribute accuracy.

The method is tested on three case studies. Data matching proves to be quite efficient (error levels less than 4%), which lead to more accurate quality results. By matching corresponding objects, the data completeness approach is able to find objects that are missing from the VGI dataset (data omission), as well as objects that are present in the VGI but not in the reference dataset (data commission or VGI over-completeness), which also broadens the method usage for data fusion purposes.

Acknowledgments

The completion of this thesis can be attributed to three major groups or persons, who I would like to acknowledge.

Firstly I would like to thank the Geographical Branch and Hellenic Military Geographical Service for supporting my nomination, and the Hellenic Army General Staff for sponsoring the first two years of my research. Without them providing the necessary educational leave of absence and funding, this thesis would have been impossible, even as a thought.

Many thanks go to my first and second supervisors, Dr. Mordechai (Muki) Haklay and Dr. Claire Ellul respectively. They provided valuable feedback, support and significantly aimed me to frame my work and complete it within reasonable time. They were always close when needed, showing such levels of professionalism and information channels between student and supervisors that I've not met before. I feel very lucky and honoured being supervised by them.

Finally but most importantly, I would like to thank my wife Christina, who unwillingly followed me to a foreign country along with our new-born daughter, tolerated my absence from home in order to study, and showed extreme patience when, after our repatriation and while we both were fully employed, I still had to find the time to complete the research.

This thesis is dedicated to my daughter, Iole – Vassileia, with the hope that one day she will be able to read it, but most importantly, that it will support my efforts to offer her and my wife a better future.

Table of Abbreviations and Acronyms

Acronym	Definition
AJAX	Asynchronous Java Script and XML
API	Application Programming Interface
BNG	British National Grid (Reference System)
BOS	Buffer-Overlay-Statistics
CC-by-SA	Creative Commons by Share Alike
CRSSA	Center for Remote Sensing and Spatial Analysis
EPSRC	Engineering and Physical Sciences Research Council
FGDC	Federal Geographic Data Committee
CSV	Comma Separated Values (file format)
GB	Great Britain
GI	Geographic Information
GIS	Geographic Information Systems
GPS	Global Positioning System
GMM	Google Map Maker
GPX	GPS eXchange
HMGS	Hellenic Military Geographical Service
IBM	Increasing Buffer Method
ISO	International Standards Organisation
ITN	Integrated Transport Network
KML	Keyhole Mark-up Language
NMA	National Mapping Agency
MO	Mapping Organisation
ODbL	Open Database License
OGC	Open Geospatial Consortium
OS	Ordnance Survey
OSM	OpenStreetMap
PHP	PHP: Hypertext Preprocessor
POI	Point of Interest
PPGIS	Public Participatory GIS
RMSE	Root Mean Square Error
SDTS	Spatial Data Transfer Standards
UN	United Nations (in this thesis it is also used for the ‘MINUSTAH’ dataset)
USGS	United States Geological Survey
VGI	Volunteered Geographic Information
XML	Extensible Markup Language

Table of Contents

Table of Figures 13
Table of Tables 18
1. Introduction 22
1.1. The rapid increase of free spatial data on the web 22
1.2. The status-quo in Cartography and Mapping 23
1.3. The need for a change 24
1.4. The geography of non-experts 25
1.5. Potential users of crowd-sourced information 26
1.6. Choosing between official and crowd-sourced data 28
1.7. Spatial Data Quality and VGI 29
1.8. Motivation 31
1.9. Research aim and questions 33
1.10. The Contributions of this thesis to VGI research 34
1.11. Outline of the thesis 36
2. Volunteered Geographic Information (VGI) 38
2.1. Introduction 38
2.2. The emergence and definitions of VGI 38
2.3. Some examples of crowd-sourced or VGI projects 43
2.3.1. VGI Source: OpenStreetMap (OSM) 46
2.4. VGI Characteristics 51
2.4.1. Perspectives of VGI 51
2.4.2. Data quality of VGI 53
2.4.3. VGI quality standards, metadata and heterogeneity 55
2.4.4. Credibility of VGI 57
2.5. Quality issues of VGI raw data 60
2.5.1. GPS technology limitations 60
2.5.2. Attribute incompleteness 61
2.5.3. Satellite imagery accuracy 61
2.5.4. Digitisation in VGI 62
2.5.5. Combination of error sources 62
2.6. Summary 62
3. Spatial Data Quality in Geographic Information Science 65

3.1. Introduction 65
3.2. Elements of Spatial Data Quality 66
3.3. Establishment of metrics to measure data quality elements 71
3.3.1. The necessity for data matching 72
3.3.2. Assessing data completeness 76
3.3.3. Assessing attribute accuracy 77
3.3.4. Assessing positional accuracy of point and linear features 79
3.4. Previous research on VGI quality 83
3.5. Summary 88
4. Methodology 91
4.1. Introduction 91
4.2. Gaps in the Literature 91
4.3. Research Objectives 92
4.4. Methodology overview 93
4.5. Data preparation 95
4.6. Reference and VGI clipping 95
4.7. Tile classification 96
4.8. Data matching (7 stages) 97
4.8.1. Data matching overview 97
4.8.2. Feature segmentation 97
4.8.3. Stage 1 98
4.8.4. Stage 2 102
4.8.5. Stage 3 102
4.8.6. Stage 4 103
4.8.7. Stage 5 104
4.8.8. Stage 6 105
4.8.9. Stage 7 106
4.9. Resulting datasets 108
4.10. Data completeness (VGI omission and commission) 108
4.11. Attribute accuracy 111
4.11.1. General 111
4.11.2. Method description 112
4.11.3. Attribute accuracy assessment 113
4.12. Positional Accuracy 115

4.12.1. General 115
4.12.2. The binary search algorithm 116
4.12.3. Positional accuracy assessment 118
4.12.4. Defining outliers 119
4.13. Ending the process 120
4.13.1. Exported information 120
4.13.2. VGI commission: indication from road type correspondence 122
4.13.3. VGI commission: indication from attribute accuracy 123
4.13.4. VGI commission: indication from tile completeness results 123
4.14. Data matching evaluation, errors and impact on quality results 123
4.15. Justification of parameters and options 126
4.15.1. Tile size and shape 127
4.15.2. Search distance parameters a,c,w 128
4.16. Areas and datasets for the three case studies 129
4.17. Summary 130
5. First case study: Urban and Rural area 132
5.1. Introduction 132
5.1.1. Reference Data Sources: Ordnance Survey's MasterMap - ITN Layer 132
5.2. Area justification and data preparation 133
5.3. Method application and results 135
5.4. Evaluation 147
5.4.1. Contribution of stages 147
5.4.2. Object matching efficiency 147
5.4.3. Attribute accuracy efficiency 150
5.4.4. Positional accuracy efficiency 152
5.4.5. VGI commission indication 154
5.5. Discussion 155
5.5.1. Data matching errors and quality results 155
5.5.2. Road types correspondence 156
5.5.3. VGI commissioned data 159
5.5.4. Topology 162
5.5.5. Spatial Patterns 165
5.5.6. Comparison with results from previous studies 166
5.6. Summary 168

6. Second case study: England and Wales 171
6.1. Introduction 171
6.2. Area justification and data preparation 171
6.3. Method application and results 173
6.4. Evaluation 189
6.4.1. Object matching efficiency 189
6.4.2. Attribute accuracy efficiency 192
6.4.3. Positional accuracy efficiency 193
6.4.4. VGI commission indication 195
6.5. Discussion 196
6.5.1. Data matching errors and quality results 196
6.5.2. Correlation of quality results: spatial patterns 196
6.5.3. VGI commissioned data 200
6.5.4. Topology 203
6.5.5. Comparison with results from previous studies 204
6.6. Summary 205
7. Third case study: Haiti (Port-au-Prince) 208
7.1. Introduction 208
7.1.1. Reference Data Sources: United Nations' 'MINUSTAH' dataset for Haiti 208
7.2. Area justification and data preparation 210
7.3. Results 213
7.3.1. UN (reference) and GMM (VGI) comparison 213
7.3.2. GMM (reference) and OSM (VGI) comparison 216
7.4. Evaluation 219
7.4.1. Contribution of stages 219
7.4.2. Object matching efficiency 220
7.4.3. Attribute accuracy efficiency 221
7.4.4. Positional accuracy efficiency 222
7.4.5. VGI commission indication 223
7.5. Discussion 223
7.5.1. Data matching errors and quality results 223
7.5.2. Road types correspondence 224
7.5.3. VGI commissioned data 226
7.5.4. Topology correction 230

7.6. Summary 232
8. Discussion 235
8.1. Introduction 235
8.2. VGI heterogeneity and tiling approach 235
8.2.1. Tile selection 236
8.2.2. The use of extended tiles 238
8.2.3. Tile classification 242
8.3. Data matching and stages order 245
8.4. Selecting target percentage for positional accuracy 246
8.5. Other issues related to the performance of the method 249
8.5.1. Automation and necessary manual intervention 250
8.5.2. Performance in computational terms 251
8.6. Implications on VGI 252
8.6.1. VGI and standardization 252
8.6.2. VGI and official data sources 253
8.7. Limitations of the method 254
8.8. Potential usage of this research 258
8.8.1. National Mapping Agencies (NMAs) already using crowd-sourced information 258
8.8.2. Commercial Mapping Organisations (MOs) already using crowd-sourced information 260
8.8.3. Opportunities for NMAs and other commercial MOs 260
8.8.4. Disaster management 261
8.8.5. Governmental, non-governmental organizations and VGI projects 262
8.8.6. Defense mapping 263
8.9. Summary 264
9. Conclusion 266
9.1. Introduction 266
9.2. Meeting the Research Aim and Objectives 266
9.2.1. Understand the nature of VGI linear data 266
9.2.2. Develop a suitable automated data matching procedure 267
9.2.3. Perform quality analysis 267
9.3. Final conclusions and opportunities 269
9.4. Suggested further research 270
9.4.1. New directions 270
9.4.2. Future improvements of the framework 272

References 276
APPENDIX A: Description of the developed application 291
APPENDIX B: Other characteristics of VGI 301
APPENDIX C: VGI Commission and other data matching examples 303
Epilogue 308

Table of Figures

Figure 2.1: Google Map Maker data availability (Source: Google Map Maker, 2010)..... **44**

Figure 2.2: OpenStreetMap webpage (www.openstreetmap.org)..... **47**

Figure 3.1: Comparison of accuracy and precision (from Servigne et al., 2006, p.184)..... **69**

Figure 3.2: Comparing thematic attributes..... **78**

Figure 3.3: Case of distortion of individual points (from van Niel and McVicar, 2002, p.459)..... **80**

Figure 3.4: Increasing buffer method (from Goodchild and Hunter,1997, p.301)..... **81**

Figure 3.5: Polygon areas after applying buffer in both lines (from Tveite and Langaas, 1999, p.33) **82**

Figure 4.1: Flow diagram of the developed methodology (with section index in parenthesis)..... **94**

Figure 4.2: Flow diagram of the data matching process..... **97**

Figure 4.3a: Possible scenarios and b: worst case scenario for the calculation of angular tolerance..... **100**

Figure 4.4: Results from actual and simplified equation for angular tolerance..... **100**

Figure 4.5: Directional segment matching: Smaller segments demand bigger angular tolerances.. **101**

Figure 4.6: Data matching challenges in stage 4..... **103**

Figure 4.7a: Datasets before the matching process, b: Matched segments before stage 5, c: Matched features after stage 5..... **104**

Figure 4.8: Matching errors before stage 7..... **106**

Figure 4.9a: Reference & VGI datasets, b: Reference matching percentages (VGI completeness), c: VGI matching percentages, d: Mixed matching percentages..... **109**

Figure 4.10: The binary search algorithm..... **117**

Figure 4.11: Case of matched objects that demand a higher buffer value if extended tiles are not used..... **119**

Figure 4.12: Problems with corresponding roads when using administrative boundaries as tiles... **128**

Figure 5.1: Areas, datasets and tiles for the 1st case study: a.Urban area, b.Rural area..... **134**

Figure 5.2: Urban area: Matched reference (ITN) dataset..... **135**

Figure 5.3: Urban area: Matched VGI (OSM) dataset..... **136**

Figure 5.4: Urban area: Non-matched VGI (OSM - red) and reference (ITN - green) datasets..... **136**

Figure 5.5: Rural area: Matched reference (ITN) dataset..... **137**

Figure 5.6: Rural area: Matched VGI (OSM) dataset..... **137**

Figure 5.7: Rural area: Non-matched VGI (red) and reference (green) datasets..... **138**

Figure 5.8: Urban area – data completeness and positional accuracy, a: ITN matching percentages (OSM completeness) b: OSM matching percentages (OSM commission) c: Mixed percentages (level of agreement between datasets) d: Positional accuracy..... **139**
Figure 5.9: Urban area – primary name, a: ITN percentages (OSM attribute accuracy) b: OSM percentages..... **140**
Figure 5.10: Urban area – secondary name, a: ITN percentages (OSM attribute accuracy) b: OSM percentages..... **140**
Figure 5.11: Urban area – total names attribute accuracy, a: ITN percentages (OSM attribute accuracy) b: OSM percentages..... **141**
Figure 5.12: Rural area – data completeness and positional accuracy, a: ITN matching percentages (OSM completeness) b: OSM matching percentages (OSM commission) c: Mixed percentages (level of agreement between datasets) d: Positional accuracy..... **141**
Figure 5.13: Rural area – primary name, a: ITN percentages (OSM attribute accuracy) b: OSM percentages..... **143**
Figure 5.14: Rural area – secondary name, a: ITN percentages (OSM attribute accuracy) b: OSM percentages..... **143**
Figure 5.15: Urban area – total names attribute accuracy, a: ITN percentages (OSM attribute accuracy) b: OSM percentages..... **144**
Figure 5.16: Matching percentages' distribution for urban and rural areas..... **145**
Figure 5.17: Use of Skewness for data distribution..... **146**
Figure 5.18: Randomly selected tiles for manual evaluation..... **148**
Figure 5.19: Example of manually evaluating feature matching..... **148**
Figure 5.20: Failure in matching corresponding objects in rural areas..... **149**
Figure 5.21: Matching error compared to number of tiles evaluated..... **150**
Figure 5.22: Manual evaluation of attribute accuracy measurement..... **151**
Figure 5.23: Positional accuracy and a: Data matching errors, b: Different representation..... **153**
Figure 5.24: Positional accuracy and big distance between corresponding objects..... **153**
Figure 5.25a: Data matching error (Primary road corresponding to ITN's A Road, not matched), b: OSM inconsistency: Rectangle defined as primary road..... **162**
Figure 5.26: Rural area, OSM commission..... **163**
Figure 5.27: Urban area, OSM commission (image rotated), tile TQ4281..... **163**
Figure 5.28: Nothern London: Richer OSM dataset (area 1 of Figure 5.8b)..... **165**
Figure 5.29: OSM and OS's Meridian 2 comparisons across five areas in London (from Haklay, 2010c, p.696)..... **167**

Figure 6.1: Regions examined in the 2nd case study..... **172**

Figure 6.2: Example of region border tiles and included data, processed individually for each region..... **173**

Figure 6.3: Data Completeness: ITN matched percentages (VGI completeness compared to reference)..... **175**

Figure 6.4: Data Completeness: OSM matched percentages (Red indicate VGI commission)..... **176**

Figure 6.5: Data Completeness: Level of data agreement (average of ITN and OSM percentages).. **177**

Figure 6.6: OSM Positional accuracy..... **178**

Figure 6.7: ITN Primary name percentages (OSM primary name accuracy)..... **179**

Figure 6.8: OSM Primary name percentages..... **180**

Figure 6.9: ITN Secondary name percentages..... **181**

Figure 6.10: OSM Secondary name percentages..... **182**

Figure 6.11: ITN Total attributes accuracy (Primary and secondary name percentages)..... **183**

Figure 6.12: OSM Total attributes accuracy (Primary and secondary name percentages)..... **184**

Figure 6.13: Tiles manually evaluated..... **190**

Figure 6.14: Data matching error compared to number of tiles evaluated..... **192**

Figure 6.15: Tiles manually evaluated for attribute accuracy..... **193**

Figure 6.16: Tiles considered as outliers for positional accuracy and evaluated area..... **194**

Figure 6.17: OSM (red) and ITN (yellow) matched data in London area: OSM dislocation to the south, (Satellite imagery source: Bing Maps provided by ESRI's ArcMap)..... **198**

Figure 6.18: Severn: Secondary names examination finds systematic errors in OSM tagging (red: OSM, green:ITN) (area 1 of Figure 6.10) **199**

Figure 6.19: Isle of Wight: Possible VGI commission regarding secondary road names (red:OSM, green:ITN) (area 2 of Figure 6.10) **200**

Figure 6.20: OSM commission (upper red feature) and data matching error (lower red feature..... **201**

Figure 6.21: OSM supremacy in Lake District area (paths, footways, tracks, etc)..... **201**

Figure 6.22: OSM commissioned data in a new built-up area (Satellite imagery source: Bing Maps provided by ESRI's ArcMap)..... **202**

Figure 6.23: Incorrect VGI topology and its effects on data matching..... **203**

Figure 6.24: Length difference between OSM and OS's Meridian 2 datasets. Black= areas of good OSM coverage; grey=areas of poor OSM coverage (from Haklay, 2010c, p.693) **204**

Figure 7.1: Incompatibility between UN datasets in the area of Port-au-Prince, **a:** 'hti_rdsl1' (major roads) and 'hti_rdsl2' (major and minor roads) **b:** 'hti_rdsl1' and 'portauprince_rdsl'.....**209**

Figure 7.2: Area and datasets studied in the 3rd case study..... **212**

Figure 7.3: Matched reference (UN - green) and VGI (GMM - red) dataset..... **213**
Figure 7.4: Non-matched reference (UN - green) and VGI (GMM - red) dataset..... **214**
Figure 7.5: Haiti area – data completeness and positional accuracy, a: UN matching percentages (GMM completeness) b: GMM matching percentages (GMM commission) c: Mixed percentages (level of agreement between datasets) d: Positional accuracy..... **215**
Figure 7.6: Haiti area – primary name, a: UN percentages (GMM attribute accuracy) b: GMM percentages..... **215**
Figure 7.7: Matched reference (GM) and VGI (OSM) dataset..... **217**
Figure 7.8: Non-matched reference (GM) and VGI (OSM) dataset..... **217**
Figure 7.9: Haiti area – data completeness and positional accuracy, a: GMM matching percentages (OSM completeness) b: OSM matching percentages (OSM commission) c: Mixed percentages (level of agreement between datasets) d: Positional accuracy..... **218**
Figure 7.10: Haiti area – primary name, a: GMM percentages (OSM attribute accuracy) b: OSM percentages..... **218**
Figure 7.11: Randomly selected tiles for manual evaluation..... **220**
Figure 7.12: Data matching error levels between a: UN and GMM, b: GMM and OSM datasets..... **221**
Figure 7.13: Different representation of real world objects between datasets..... **222**
Figure 7.14: UN-GMM datasets: Failed data matching due to distance. Mislocated reference dataset?..... **228**
Figure 7.15: U.N. dataset with **a:** 2003 Google maps image, **b:** 2010 Google maps image, **c:** 2010 Google maps image and GMM non-matched (commissioned) data (in red)..... **229**
Figure 7.16: GMM-OSM datasets: OSM commission example..... **230**
Figure 7.17: Topological errors, a and b: Features not divided at road intersections, c: Use of multi-linestring..... **231**
Figure 7.18: Data matching error levels between topologically corrected datasets and original ones..... **232**
Figure 8.1: Variations in data comparison due to data splitting..... **238**
Figure 8.2: VGI and reference segment length frequency of first case study..... **239**
Figure 8.3: Benefits in data matching when using the extended tile..... **240**
Figure 8.4: Benefits in positional accuracy when using the extended tile..... **241**
Figure 8.5: Manually examined cells where by applying rural only constraints, data matching was fixed or deteriorated..... **243**
Figure 8.6: Urban (left) and rural (right) areas tested to decide on target percentage..... **247**
Figure 8.7: Buffer widths corresponding to 1% increment of target percentage (90-100%)..... **248**

Figure 8.8: Average and median buffer increment corresponding to 1% increment of target percentage..... **248**
Figure 8.9: Errors in data matching due to insufficient geometric constraints of stage 1..... **254**
Figure A-1: First page: Providing the user credentials..... **289**
Figure A-2: Defining datasets..... **290**
Figure A-3: Defining tile names..... **290**
Figure A-4: Checking datasets consistency..... **291**
Figure A-5: Selecting road types..... **292**
Figure A-6: Details on reference and VGI road types..... **292**
Figure A-7: Customisation of the positional accuracy approach..... **293**
Figure A-8: Notifications of the comparison progress..... **294**
Figure A-9: Visualisation of the progress using QGIS, a: currently processing TQ0682, b: data refresh after 5 seconds..... **298**
Figure C-1: London region, OSM commission (parking area at Heathrow airport)..... **301**
Figure C-2: London region, OSM commission..... **301**
Figure C-3: Lancashire region, OSM commission..... **302**
Figure C-4: Severn region: OSM seems to have inconsistent data..... **302**
Figure C-5: Humberside region, OSM commission..... **302**
Figure C-6: Essex region, OSM commission..... **303**
Figure C-7: Yorkshire region, OSM commission..... **303**
Figure C-8: North region, routes for pedestrians – bicycles..... **303**
Figure C-9: Haiti area, UN-GMM datasets: GMM commission, example 1..... **304**
Figure C-10: Haiti area, UN-GMM datasets: GMM commission, example 2..... **304**
Figure C-11: Haiti area, UN-GMM datasets: Failed data matching due to distance. Mislocated reference dataset?..... **304**
Figure C-12: Haiti area, UN-GMM datasets: GMM commission, example 3: Non-traffic road type. **305**
Figure C-13: Haiti area, GMM-OSM datasets: OSM commission, example 1..... **305**
Figure C-14: Haiti area, GMM-OSM datasets: OSM commission, example 2..... **305**

Table of Tables

Table 1.1: Official and crowd-sourced spatial information: considerations regarding their usage..... **29**
Table 4.1: String similarity scores for various cases..... **103**
Table 4.2: General cases of matching for each cell..... **108**
Table 4.3: Reference road type correspondence..... **110**
Table 4.4: Examples of the binary search algorithm (target percentage 95%) for 4 tiles **118**
Table 4.5: Example of the exported quality information (length table - no positional accuracy)..... **121**
Table 4.6: Example of the exported positional accuracy information..... **121**
Table 4.7: Applied values for the parameters used during data matching approach..... **126**
Table 4.8: Applied values for the parameters used during attribute and positional accuracy.....**127**
Table 5.1: Studied areas and road network information..... **133**
Table 5.2: Resulting network lengths for study areas..... **138**
Table 5.3: Statistics for the urban area (Greater London)..... **146**
Table 5.4: Statistics for the rural area (West of Newcastle)..... **146**
Table 5.5: Contribution of stages to data matching..... **147**
Table 5.6: Data matching errors..... **149**
Table 5.7: Attribute accuracy errors for the urban area..... **152**
Table 5.8: Attribute accuracy errors for the rural area..... **152**
Table 5.9: Evaluation of positional accuracy outliers..... **154**
Table 5.10: Evaluation of VGI commission indication..... **155**
Table 5.11: Estimation of errors in quality results for the provided method..... **156**
Table 5.12: ITN road types correspondence..... **157**
Table 5.13: OSM road types correspondence..... **158**
Table 5.14: ITN matched road types: what is mapped by OSM and what is not..... **159**
Table 5.15: OSM matched road types: what is mapped by ITN and what is not..... **160**
Table 5.16: Results for rural area with corrected topology..... **164**
Table 5.17: Positional accuracy results compared to Haklay's (2010c) **166**
Table 5.18: OSM completeness provided by this study for the five London areas examined by Haklay (2010c) **168**
Table 6.1: Studied regions and road network information..... **172**
Table 6.2: Resulting network lengths for study areas..... **174**
Table 6.3: Statistics for East Anglia region..... **185**
Table 6.4: Statistics for Essex region..... **185**

Table 6.5: Statistics for Humberside region..... **185**
Table 6.6: Statistics for Lancashire region..... **186**
Table 6.7: Statistics for Manchester region..... **186**
Table 6.8: Statistics for Midlands region..... **186**
Table 6.9: Statistics for North region..... **187**
Table 6.10: Statistics for Severn region..... **187**
Table 6.11: Statistics for South region..... **187**
Table 6.12: Statistics for South East region..... **188**
Table 6.13: Statistics for South West region..... **188**
Table 6.14: Statistics for Wales..... **188**
Table 6.15: Statistics for West region..... **189**
Table 6.16: Statistics for Yorkshire region..... **189**
Table 6.17: Data matching errors (per region – dataset and total)..... **190**
Table 6.18: Attribute accuracy errors..... **193**
Table 6.19: Outliers found in each region..... **194**
Table 6.20: Outliers manually examined..... **195**
Table 6.21: Evaluation of VGI commission indication..... **196**
Table 6.22: Estimation of errors in quality results for the provided method..... **196**
Table 6.23: Average quality values per region, highlighting highest (green) and lowest (red) scores..... **197**
Table 6.24: Tiles with OSM commission..... **202**
Table 7.1: Studied areas and road network information for 3rd case study..... **213**
Table 7.2: Resulting UN and GMM network lengths for Haiti area..... **214**
Table 7.3: Statistics for Haiti area (UN and GMM datasets)..... **216**
Table 7.4: Resulting GMM and OSM network lengths for Haiti area..... **216**
Table 7.5: Statistics for Haiti area (GMM and OSM datasets)..... **219**
Table 7.6: Contribution of stages to data matching for the two Haiti cases (UN-GMM and GMM-OSM)..... **219**
Table 7.7: Data matching errors between UN and GMM datasets..... **221**
Table 7.8: Attribute accuracy errors..... **221**
Table 7.9: Text similarity failure and success (the latter underlined in bold-italic) in matching strings..... **222**
Table 7.10: Evaluation of positional accuracy outliers..... **223**
Table 7.11: Evaluation of VGI commission indication..... **223**

Table 7.12: Estimation of errors in quality results for the provided method..... **224**

Table 7.13: UN and GMM road types correspondence..... **224**

Table 7.14: GMM and OSM road types correspondence..... **225**

Table 7.15: UN and GMM road types: what is mapped by the other dataset and what is not..... **226**

Table 7.16: GMM and OSM road types: what is mapped by the other dataset and what is not..... **227**

Table 7.17: Data matching errors between UN and GMM datasets topologically non-corrected.... **232**

Table 8.1: Differences for 35 tiles in central London because of tile shifting: quality results minus quality results for shifted tiles **237**

Table 8.2: Compared and matched lengths in rural area for different splitting methods..... **238**

Table 8.3: Compared and matched lengths in urban area for different splitting methods..... **238**

Table 8.4: VGI and reference segment length statistics for the first case study..... **240**

Table 8.5: Benefits in quality evaluation when using the extended tile..... **241**

Table 8.6: Statistics on total length and number of features and junctions when rural only constraints are applied in the rural area of first case study..... **243**

Table 8.7: Tile classification of the first case study..... **244**

Table 8.8: Matching percentages with and without tile classification..... **244**

Table 8.9: VGI and reference segment length statistics (m) for the first case study **249**

Chapter 1

Introduction

1. Introduction

1.1. The rapid increase of free spatial data on the web

Web 2.0 is the evolution of web that allows interactivity between users and web pages. Its technologies are changing our lives in ways no one could imagine only a decade ago. Online collaboration is now a fact; it is relatively easy for someone to be part of online social networks, to begin or participate in a discussion, to be heard and influence other's opinions. Additionally, one can add or alter information and edit online databases, contributing in publicly accessible knowledge sources. Examples are personal blogs, web sites with customers' opinions on the quality of a product, (e.g. Amazon (2011) for shopping, Tripadvisor (2010) or Booking (2010) for travelling) and applications that collect other types of information, like the free online encyclopaedia 'Wikipedia' (2010). The unknown, yet significant contributor can be anyone, from a professional who expresses an opinion based on his or her expertise, to an unskilled person, yet with adequate computer skills.

One of the areas affected by Web 2.0 is Cartography and Mapping in general. Answering the question *'Why is the Web an interesting medium to present and disseminate geospatial data?'*, Kraak and Brown (2001, p.2) explain that *'information on the web is virtually platform-independent, unrivalled in its capacity to reach many users at minimal costs and easy to update frequently. Furthermore and more particularly in relation to maps, it allows for a dynamic and interactive dissemination of geospatial data, offering new mapping techniques and use possibilities not seen before with traditional printed maps, such as multimedia integration"*. As a result, many offer tools for simple users to produce and share Geographic Information (GI) on the internet, including also commercial vendors; Google Maps, Google Earth, Common Census, WikiMapia, OpenStreetMap, Microsoft Virtual Earth, Yahoo! Maps and The Open Planning Project (McConchie, 2008). By combining and enhancing the tools of Web 2.0 (discussed in section 2.2), the user is now able to avoid the cost of purchasing digital GI produced by a National Mapping Agency (NMA) or other commercial providers, and to find GI that will cover specific demands for non-traditional spatial products (e.g. digital maps for cycling or skiing). The non-expert user (in terms of mapping or cartography) can now rely on the web and by using the provided Application Programming Interfaces (APIs)[1], one can produce customised maps or other spatial derivatives. Although the use of APIs demands a certain level of computer skills, it can be done with virtually no cost (a computer

[1] API: '*A language and message format used by an application program to communicate with the operating system or some other control program such as a database management system (DBMS) or communications protocol*' (PC Magazine, 2012).

and an internet connection is needed) and, furthermore, the final product can be disseminated and used by other users as well. An example is OpenStreetMap (OSM), a project started in 2004 from the UK as a result of the high prices (back then) of data provided by the NMA of Great Britain, the Ordnance Survey (OS). Relying on users who volunteer to gather spatial data, OSM started creating vector maps all over the world. Today, as the number of its users is growing more and more (OpenStreetMap, 2010a), so does the world coverage, and maps can be created and downloaded.

This rapid increase of spatial data on the web has attracted the attention of researchers and many terms are already given to this new trend and its procedures, such as 'Crowdsourcing' (Howe, 2006), 'Neogeography' (Turner, 2006), 'Volunteered Geographic Information' (Goodchild, 2007b), 'User Generated Spatial Content' (Antoniou *et al.*, 2010b). These terms are further analysed in section 2.2. Questions arise, such as the reasons of participation, the digital exploitation, the digital divide or copyright issues, but there can also be more technical challenges, referring to the way the data are gathered, its positional accuracy, its completeness or its richness in attributes.

1.2. The status-quo in Cartography and Mapping

Cartography and mapping in general has so far been an area restricted to professional cartographers, who were usually considered to have an unquestionable expertise. NMAs have traditionally been the official source of spatial data, which they produce and trade. Advances in technology and market demands enabled private companies to also become providers of spatial data, using specialised personnel too. These private spatial providers will be referred to as Mapping Organisations (MOs). These MOs may not produce the broad range of spatial products that an NMA does – or may be obliged to, according to legislation and institutional regulations that force an NMA to cover the whole country. MOs will chose to produce specific spatial products, which are in demand and profitable, usually avoiding products that cover rural and remote areas or the needs of relatively small market groups. There are cases where MOs produce data of higher quality for a specific purpose or area than an NMA. Such a case in the UK is UKMap (2011), which produces spatial data based on 1:1,000 scale topographic mapping (UKMap, 2011) as compared to 1:1,250-1:10,000 base mapping scale of GB's NMA, (Ordnance Survey, 2012), however, so far it is restricted to London area only. While both NMAs and MOs use their own standards and can provide consistent data, there are differences in the way they are funded, their organisational structure, their flexibility to adapt to market demands and their authority. NMAs remain the formal source of spatial data; a source that one can rely on to strengthen a decision-making procedure or at court for disputes regarding property boundaries.

The fact that non-experts gather spatial data and create geographical information, which traditionally is the job of delegated government agencies or commercial vendors, creates the necessity to understand why this started to happen. Next, it is essential to understand how to address this new source; is it ephemeral or something more serious? Can it substitute for official data, is it something that complements it or is it something totally different that cannot be combined with it?

1.3. The need for a change

Goodchild (2007a) argues that maps are not constantly being updated or made more accurate. The reduction in government funding, partly as a result of a global economic recession, hinders updating even for the technologically leading countries, including the U.S. (Goodchild, 2007a), putting NMAs' authority under pressure and leading many customers to lose trust in them.

Another reduction in funding stems from a gradually and obligatory need to release their spatial data to the public for free, as a result of a constant pressure for a free access to information funded by taxpayers. Directive 2003/98/EC (Europa Information Society, 2010) and the Freedom of Information Act 2000 (OPSI, 2010) prove that legislation is changing to provide this freedom. Switzerland's 'Swisstopo' is an example of an NMA offering free data access 'in order to promote their use' in 1 January 2010, following (or forced by) a new Ordinance (Swisstopo, 2010). A second example is Great Britain's NMA, the Ordnance Survey, who released a series of products since 1st of April 2010 (BBC news, 2010). As a third example, Finland's NMA, the National Land Survey, opened its topographic datasets to the public on 1 May 2012 (NLS, 2012). However, by freeing all or some of their data, NMAs potentially lose financial sources that could otherwise be used for updating their spatial products.

Data provided by a NMA or MO are typically accompanied by a quality assurance procedure, which guarantees the quality of the source. This procedure refers to all the necessary stages to produce spatial information, starting from data collection. Raw data are processed using a specified sequence of steps, and the final product is usually evaluated by sampling. This is a strenuous series of tasks that takes time, increases the cost of the product, reduces the production or frequency of updates and will not always be understood or needed by the user (Boin and Hunter, 2007; Coote and Rackham, 2008). It also makes it difficult to change the line of production based on the market demands, because this would probably need a different quality assurance procedure.

Advances in technology and access to higher education provides people with a better understanding of space. Geographic terms become more understandable to non-geographers. This helps them raise questions on how fit-for-purpose the official source is, especially when there is a relatively high cost to acquire data. On the other hand, research shows that users usually ignore quality information (Agumya and Hunter, 1999; Devillers *et al.*, 2005; van Oort, 2006), which is among the major advantages of official and professional sources. Adding to this, demands for different and non-traditional spatial products urge for different types of spatial sources, as they cannot be fully satisfied by the existing organisations.

1.4. The geography of non-experts

The current state of affairs is that many people are able to produce and share spatial information with practically no limits on the type of products, exploiting the new web 2.0 technology regardless of their level of expertise in geography. Provided that one has an internet connection and basic computer skills, web mapping applications that collect and share spatial information can be 'attractive' even for professionals (Brotzman, 2009). After all, some types of information do not need experts to be collected (Goodchild, 2008b). The more difficult cartographic and visualisation options are designed and pre-decided for the majority of the less qualified users (in terms of cartography), so their interface is usually more user-friendly. (e.g. the OSM user cannot change the symbols, their size and colors that are used to represent the real-world objects on the screen). Although from a geographer's point of view this deprives some freedom, it succeeds in accessing a broader audience.

The spatial attribute becomes more and more important in web applications for various purposes, only restricted by the developer's imagination. Photo sharing programs enable geo-referencing of the photos, so a simple image uploaded by an unknown user acquires a spatial meaning by being placed on a map (e.g. Flickr). Web 'mashups' (discussed in section 2.2) that combine web pages to create a new service usually include one geospatial application as background (e.g. Tripadvisor places hotels on a Google map). Other web mapping projects rely on volunteers to collect and upload GPS tracks in order to create a map that will include also non-traditional information (e.g. the road network of OSM includes footpaths, cycleways, steps). This information is usually free and combines various types of raw data for a specific purpose. However, as free information, it can also be further combined with other types of information for purposes much different than the ones it was originally designed. Depending on the number of dedicated users, it can be frequently updated, resulting in a rapidly growing coverage.

Spatial data on the web can be lines (e.g. road network), points (e.g. points of interest, location of a geo-referenced photo) or polygons (e.g. administrative boundaries). According to Doytsher *et al.* (2001) and Ramirez and Ali (2003), the majority of map features in general are linear. The road network is considered among the most important spatial information, because human activities usually depend on it. When a built-up area is expanded due to a population increase, the road network is the backbone of the new area development. Routing applications use the road network and their quality as a service depends on how complete and updated the road network is. As a result, it is the primary data type collected by many 'crowd-sourced' applications (applications that collect information through anonymous users), such as OSM or Google Map Maker.

1.5. Potential users of crowd-sourced information

One next question is who might be interested in this type of information and how it could be used. The simplest form of interested users could be uncategorised average users for various personal reasons, such as vacation planning, hiking, navigation through a GPS unit, finding places and other points of interest, finding geo-referenced pictures of the area to get a first impression before even arriving there. This, however, is a less demanding type of usage, and wrong usage will unlikely cause any risks to others. Parker *et al.* (2010) define other types of stakeholders, namely 'Special Interest Mapping Groups', 'Local Communities' and 'Professionals'. A different distinction could be based on the level of interaction with the data, dividing the users to 'viewers' (who access and use the data) and 'contributors' (who additionally enrich or modify the data).

A more advanced usage of crowd-sourced data is to update national spatial databases. Goodchild (2007a) supports that this could be achieved by 'using citizens as voluntary sensors'. Budhathoki *et al.* (2008, p.156) view Volunteered Geographic Information (VGI) as patchworks to SDI, provided that the role of user is redefined. Seeger (2008) also mentions that professionals could enhance their datasets by using VGI. Although he focuses on landscape planning and site design processes, he concludes noting that for a broader use and implementation of VGI, research in the areas of data quality and copyright issues (referring to ways of sharing the collected data for uses beyond its initial purpose) is necessary. The need for NMAs to include VGI into their updating process is also the conclusion of the 1st EuroSDR Workshop, which focused in crowdsourcing for updating National Databases (EuroSDR, 2009). A few NMAs are already leading the way, having already started to use volunteers for creating or updating some of their datasets. Some examples include United States Geological Survey (USGS) (Bearden, 2007), Switzerland's NMA (Guélat, 2009), GB's NMA (Ordnance Survey, 2010a; People's Place Names, 2010), and will be further discussed in section 8.9.1. Another

option, suggested by Antoniou *et al.* (2010b), is to exploit VGI information that is already gathered for other purposes. They propose that Geotagged images from photo sharing websites can provide information – under certain circumstances – about areas not visible when a map is created using aerial photography, such as pavements or buildings and property borders under trees.

MOs work similarly to NMAs, but are usually self-funded. However, their operational objectives are more flexible, while NMAs will usually have to justify a change in their publicly funded line of production through a bureaucratic and lengthy procedure. As a result, it is easier for an MO to integrate new or ambiguous data sources for their traditional data. Some well-known private sector examples of investing on crowd-sourced information are Google (Helft, 2009) and TeleAtlas (Helft, 2009; Mac Gillavry, 2009), described in section 8.9.2.

Decision-makers may also need to rely on spatial data that are more up-to-date than the official ones. In case of crisis management, VGI is a good option of a fast collection of updated data that show the new status (Goodchild, 2007a; Ostermann and Spinsanti, 2011). In case of natural disasters, this is necessary to coordinate the search and rescue teams. Examples of such VGI usage are the San Diego fire in 2007 (Majchrzak, 2011), the Katrina disaster (Mullins, 2010) and Haiti earthquake (Mullins, 2010; Haklay, 2010b).

Harrison and Haklay (2002) provide examples of using Public Participation Geographic Information Science (PPGIS) in decision planning in London and USA. Although PPGIS differs from VGI, as will be explained in section 2.2, these examples show how local authorities can collect citizen information for a more interactive and collaborative planning. Seeger (2008) refers to the use of VGI by local government, state agencies or community-based organisations for landscape planning and site design process, giving the CommonCensus Project as an example.

Decision planning or studies of non-profit organisations sometimes need to access official as well as VGI sources and combine them. This combination of two data sources in order to create a new one is called conflation (further discussed in section 3.3.1). An example is EPSRC's research project on adaptable suburbs (EPSRC, 2011), which studies the relationship between networks of human activity. Another rather theoretical example is an environmental non-governmental organisation that monitors wildlife on a mountain and needs spatial information on footpaths, as well as the basic road network that cross the area.

Developing countries that strive to improve their infrastructure but are hindered by technology and budget limitations could use crowd-sourced information and open sources to facilitate their development. Iliffe (2011) provides an example in Kenya, where services of sanitation, waste management and water are mapped using community members as surveyors and open source GIS technologies. He further supports the engagement of citizens because projects become sustainable at low cost, governmental processes are more transparent while at the same time citizens realise the limitations of the government and, finally, citizens are more familiar with the problems and their geography. This crowd-sourced information could be used by the government itself, as well as other organisations in developed countries that aim to help.

1.6. Choosing between official and crowd-sourced data

The interested in spatial data, whether an individual or an organisation such as the ones described in the previous section, may now have to choose between using official or non-official spatial data, at cost or for free. Although the quantity and variability of spatial information on the web broadens the horizons of its usage, each case needs careful consideration regarding the data source and its suitability for a specific purpose. This fitness-for-purpose analysis is essential to decide whether and how to use VGI, because the use of an unsuitable spatial dataset, in terms of quality, may lead to wrong decisions or erroneous products, giving causes for misinterpretation (Kraak and Brown, 2001).

Although Volunteered Geographic Information (VGI) characteristics are described in Chapter 2 in more details, Table 1.1 provides some general ones that could be taken into consideration when choosing between NMA/MO and VGI datasets, and shows that data quality information, which is fundamental for this choice, is usually non-existent in VGI cases.

VGI quality is an issue because the spatial data source usually has no standards, no well-defined data structures, and no metadata or quality evaluation procedures. In contrast to official data from an NMA or MO, free spatial data sources are created by many anonymous users and do not have quality information. There may be no information on who created the data, with which method, with what positional accuracy, or what the motivation and credibility of the user is. However, such information is essential for choosing whether to use VGI or not. Users have different requirements, and VGI quality differs between datasets. Who could use VGI depends on its spatial data quality, which in turn depends on the existence of appropriate methods to measure it.

Considerations	NMA or other official dataset	Crowd-sourced dataset
Authority	Established, could stand on a court procedure	None: nobody can be blamed or take responsibility
Standardised processes	During collection and production, quality evaluation exists	Poor: usually no quality assurance or, at best, 'many-eyes' validation, sometimes limited moderation
Copyright legal framework	Strict, limitations with the output product usage (NMA examples: OS in GB, HMGS in Greece)	Loose, free distribution or use for commercial purposes is allowed - may depend on terms
Specifications - Metadata	Yes: usually descriptive of quality, types of information collected, update process, quality evaluation	Poor or non-existing
Positional accuracy	Provided as metadata, usually for the whole dataset	Not defined, variable in an unknown pattern
Density - coverage	Standardised and homogeneous, predefined information collected	Variable, broader range of information may be collected but not uniformly
Update process	Uniform on a regular basis, applies to the whole dataset or part of it (e.g. area, thematic layer)	Variable and unpredicted frequency: daily in some areas, seldom in others
Cost to acquire	Yes (in most cases)	No (in most cases)
Cost to produce	High (personnel and equipment is necessary)	Low (volunteers are not paid)

Table 1.1: *Official and crowd-sourced spatial information: considerations regarding their usage*

1.7. Spatial Data Quality and VGI

The technical issue of spatial data quality has a longer history than VGI. Starting back in 1970-1980s, researchers have focused on defining and distinguishing quality from uncertainty, as well as what are the quality elements that need to be measured. Different points of view and technological advances have changed the significance or added new quality elements, until these were standardised by the International Organisation for Standards (ISO) in 2002. They define five quality elements (ISO/TC 211, 2010), the measurement of which is essential to communicate spatial data quality; Completeness, Logical consistency, Positional accuracy, Temporal accuracy and Thematic accuracy.

Existing methods of data quality evaluation can be applied to spatial datasets and provide a quality insight for the whole dataset. Although they are suitable for official spatial datasets, they may not always give a representative answer when applied to VGI. This type of data is far different from the ones traditionally provided by an official or commercial source. The lack of metadata and standards, combined with the unknown motivation and user credibility, lead to heterogeneous datasets with unknown quality, for which the traditional quality measures, designed for datasets with uniform quality, cannot be applied. VGI sources include areas with full coverage and areas with scarce or no contribution (completeness); there are areas with more accurate and areas with less accurate data in terms of position (positional accuracy); some data are adequately/accurately described by attributes and some are not (thematic completeness/accuracy); information on data collection date is not always available (temporal accuracy). Topology may be erroneous or missing, no or relatively loose standards are followed and usually no quality assurance exists (logical consistency). These quality issues, mentioned as a whole or partially by many researchers (van Oort, 2006; Boin and Hunter, 2007; Sieber, 2007; Goodchild, 2007a; Flanagin and Metzger, 2008; Coote and Rackham, 2008; Haklay & Weber, 2008; Goodchild, 2008a; Auer and Zipf, 2009; Antoniou *et al.*, 2010a; Haklay, 2010c; Devillers *et al.*, 2010), create the need of methods suitable to evaluate crowd-sourced datasets.

Additionally, different data types demand different quality measurement approaches, and in some cases no specific method is standardised. For example, as will be discussed in section 3.3.4, there is no standardised method to assess positional accuracy of linear features, which adds to the complexity of quality evaluation when combined with the mentioned VGI characteristics. Yet, this data type is the one used for road networks, which is of great importance in our everyday maps, as well as for other networks (hydrological, electrical power, etc).

Quality evaluation of VGI could be achieved through a comparison with an official data source of known quality. In this way the results will directly refer to the data, regardless of the user or the other VGI characteristics. The considerations of Table 1.1 could then lead to a more balanced and unbiased choice between data sources.

Some of the previous research treated VGI in this way, using a ‘reference’ or ‘ground truth’ dataset to compare it with the VGI source (Kounadi, 2009; Haklay, 2010c; Girres and Touya, 2010; Cipeluch *et al.*, 2010; Zielstra and Zipf, 2010; Ueberschlag, 2010). However, the comparison is usually performed manually or semi-manually, which restricts the examined area to a reasonable size, or

evaluates a sample instead of the whole dataset. This hinders the repetition of the method when sources are different, updated or in a different area. Most of these studies are also focused on selected quality elements and datasets, providing a partial quality evaluation. Considering the frequency of updates in VGI, many of the previous research conclusions may already be obsolete.

1.8. Motivation

Section 1.5 discussed the potential users, mentioning also some examples of existing ones. This section moves on with some more theoretical suggestions of applications, which, along with section 1.5, reveals the motivation behind this research.

When dealing with crisis management, decision-makers and authorities need to evaluate VGI before proceeding with using such datasets. Data matching and positional accuracy is essential in such cases; for example a road reported as blocked must inform a rather obsolete official dataset so that an alternative route is selected, or efforts to restore it are targeted correctly and timely. Additionally, volunteers that risk their lives in such circumstances may not have access to any other official or updated data, so they need an authoritative answer on whether they can trust a specific VGI source. This can only be provided if there is a way to perform VGI quality analysis within a reasonable timeframe.

Communicating the results of a VGI quality analysis can be helpful also for others that cannot have access to official data to perform the analysis by themselves. Non-governmental organisations or other local bodies that cannot afford official data can be informed and decide whether to use VGI or not, depending on their objectives and requirements and the quality results for their area of interest.

Developing countries are in need of useful and reliable data. Iliffe (2011) studies mapping services in the developing world and refers to Kenya as the first country that opened its datasets for education, energy, health, water and sanitation purposes, however it is through open source GIS (specifically the OSM project) and public participation that valuable and up-to-date data are collected. For the cases of unavailable or unreliable government data, being able to evaluate VGI would aid the governments of deprived environments to improve the basic services provision, as well as to improve and update their datasets.

Defining the data missing from both datasets, as compared to each other, can be further used for conflation purposes (discussed in section 3.3.1) leading to a new and more complete dataset from the ones involved in the comparison procedure. When applied for official datasets, this will put into effect researchers' ideas of updating national spatial databases (Goodchild, 2007a; Budhathoki *et al.*, 2008; Seeger, 2008), although licensing issues need also to be considered.

Specifically for professional stakeholders, there are additional opportunities apart from the usage mentioned in section 1.5 or conflation. There will be cases that VGI will not satisfy institutional and professional GI producers because their requirements in terms of data quality, timeliness, and completeness are not flexible (Budhathoki *et al.*, 2008), or because licensing may prohibit VGI integration. In the first case, by being able to evaluate VGI, they can prove that their product is of higher quality and use the results for advertising purposes to raise their income. In the second case that VGI cannot be directly imported or integrated, it could be used indirectly to point out missing spatial data or to complement information on existing data, without actually integrating VGI as a whole. According to the importance of the missing data, the updating process could be focused on these areas or specific roads instead of following an updating policy that relies on randomly choosing large areas to examine, where a comparison with VGI sources may show that there is no recent change. In this way the updating process is more effective by being less costly and faster. By updating their data where they prove to be inferior to VGI, they increase their product value in order to be competitive with contemporary data for the areas that users seem to be mostly interested.

Finally, there are also cases of sensitive spatial data. Maps are political and can be used for propaganda (Monmonier, 1996), so an NMA needs to pay close attention to VGI sources for boundary disputes or deliberate use of informal naming for political reasons. VGI is free, and the majority of people who need spatial data for personal reasons will probably resort to VGI instead of the official or commercial datasets. VGI as a propaganda means can be effective for people who are not familiar with the area. As an example, Helft (2009) describes a boundary dispute case that occurred in Google Map Maker between Pakistani and Indian users. An NMA should be able to respond the soonest possible, either by logging in and correcting the VGI source, or by following legal actions if applicable.

From the VGI point of view, informing the community on the quality results of a VGI source could be a motivation to target their efforts to marginalised areas with reduced quality and density, leading to a less heterogeneous VGI dataset. Non-governmental organisations that rely on VGI and are more

aware of the marginalised areas (e.g. Mapping For Change, 2011) could also be helped in directing volunteered mapping efforts.

1.9. Research aim and questions

Mummidi and Krumm (2008) and Brotzman (2009) refer to VGI as an 'attractive' spatial source, while the latter adds that it gains ground against official data. Many organisations (NMAs, MOs, not-for-profit ones, local authorities – some of them will be discussed in section 8.9) as well as individuals have already started to rely on such sources, each one at different levels and purposes.

However, VGI spatial data quality is difficult to assess due to the lack of appropriate tools (Walsh, 2008; Budhathoki *et al.*, 2008; Haklay, 2010c), so its usage is still restricted to the relatively 'safe' cases of information, where quality can be improved as the number of users editing the same data increases, following Linus' Law 'given enough eyeballs, all bugs are shallow' (Raymond, 1999). Such information can be the local naming for gazetteers, desired spatial products, points of interest. For more advanced information such as the road network, contributed data need to be evaluated differently. The dynamic aspect of VGI also implies that updates can be quite frequent, and VGI sources may follow a different structure for the same area and data type. These need to be considered when designing a VGI evaluation process.

There is a need for a systematic analysis of VGI (Haklay, 2010c). This should include assessment of the appropriate quality elements for VGI that better describe its quality, considering its different nature. This analysis needs also to be automated, so that it could be easily re-applied in the future in a different area or when datasets are updated. Finally, by designing it to be applicable regardless of the data sources used, the analysis could be a useful framework for VGI quality evaluation in general.

The above form the research aim of this study, which is to provide a framework for evaluating spatial data quality of VGI linear data, such as road networks. This can be achieved by answering the following research questions:

- **How can we describe spatial data quality of VGI?**
  This includes selecting the appropriate quality elements that need to be measured for VGI. The selection should take into consideration VGI nature, e.g. lack of standards or consistency renders the 'logical consistency' quality element inappropriate.

- **How do we measure these quality elements for linear VGI?**

  Heterogeneity implies new approaches for quality assessment that could provide more than one quality values, which is likely to be more representative than a single one for the whole dataset. As a result, for each quality element a method should be designed to accept VGI heterogeneity.

- **How can the quality analysis be performed in an automated and systematic way?**

  This methodology for VGI quality needs to be designed in an automated way that will compare the two input datasets and produce results regardless of the area or datasets used. This framework will enable the future repetition of the evaluation, providing valuable results for VGI quality that could extend its potential usage.

- **What will this quality evaluation tell us about VGI?**

  This refers to how useful the results are, what else we learn about VGI, possible opportunities of the methodology to tackle other research areas as well, and if there seems to be some correlation between quality elements, e.g. if higher data completeness implies better positional accuracy.

1.10. The Contributions of this thesis to VGI research

In order to answer the research questions, a method to assess the spatial quality of VGI road network is developed in this study, which has the following characteristics:

- It compares VGI datasets with official ones of known quality, using information that can be generally found in any linear dataset. This does not restrict the method application to specific datasets, unlike previous studies on VGI quality (e.g. Ludwig *et al.*, 2010).
- It is fully automated, enabling a systematic approach that so far is missing (Haklay, 2010c) and a future application, as well as application on large areas and whole datasets, unlike several past studies on VGI that were manually performed or applied on small areas (e.g. Cipeluch *et al.*, 2010) and sampled data (e.g. Girres and Touya, 2010).
- It deals with VGI heterogeneity by providing results for smaller areas, instead of a uniform quality value for all the area and dataset studied.
- It includes an automated matching procedure to ensure that corresponding objects are compared, so that quality evaluation relies on more useful indicators, unlike several past studies on VGI quality (e.g. attribute accuracy in Ueberschlag (2010) or data completeness in Zielstra and Zipf (2010)).
- The automated data matching procedure is innovative and specifically designed for VGI, so it is much more effective than the proposed ones from previous research that assume

consistency and uniform behaviour within the datasets involved (e.g. Gabay and Doytsher, 2000; Mustière and Devogele, 2008; Safra *et al.,* 2010).

- It provides information on data missing from both datasets, which makes it a valuable tool for finding the commissioned data (excessive data in one dataset, explained further in section 3.3.2).
- Positional accuracy is directly calculated, unlike previous studies, which assumed predefined values of positional accuracy (Haklay *et al.*, 2010; Ludwig *et al.*, 2010; Ueberschlag, 2010; Al-Bakri and Fairbairn, 2010).

It needs to be noted, however, that this research does not include the most appropriate method to present the results to the user. This is a research in spatial data quality and methods to measure it. The communication of the quality results is rather simplistic through output files that can be accessed using the appropriate software. User interaction with the quality results through visualisation is a research area on its own and is outside the scope of this thesis.

The novelties of this research are:

- It uses an innovative and fully automated data matching procedure, able to process the whole datasets involved (although there is also the option to exclude some road types) and isolate the common objects quite effectively (data matching error levels below 4%). The data matching approach is specifically designed for VGI and combines geometric with thematic attributes, using VGI attributes where and when they are provided.
- It includes data completeness, positional and attributes accuracy evaluation, also fully automated and applied on the corresponding objects between datasets.
- For positional accuracy it uses an existing method (Goodchild and Hunter, 1997), however it puts into practice the researchers' suggested theoretical approach for the first time, using a binary search algorithm appropriately designed and implemented. This research is arguably the first to actually calculate the positional accuracy for a user-defined level of confidence instead of calculating the level of confidence for some user-defined/assumed positional accuracy values that are treated as steps.
- Data matching is performed at feature level, enabling the creation of sub-datasets that contain the non-corresponding objects for each dataset. Due to the data matching efficiency of this study, this can also be used as the first stage of conflation (discussed in section 3.3.1), for which so far no VGI-customised method is provided.

- During data completeness, indication at feature level is provided for VGI objects that should be present to the official dataset but are missing (VGI over-completeness or commissioned data). A manual post-processing evaluation will help find VGI over-completeness by limiting the data that need to be examined and excluding data types that are unique in each dataset.
- It is a general framework that can be applied in any case of heterogeneous datasets, including the case of comparing two VGI datasets, when no official dataset is available.

1.11. Outline of the thesis

In this first chapter an introduction of the thesis and the research context were presented. Chapters 2 and 3 form the literature review part. Chapter 2 aims to provide an understanding of VGI, its characteristics and implications. Chapter 3 discusses spatial quality, justifies the necessary quality elements to be evaluated for VGI, argues on the existing measuring methods and their suitability for VGI and concludes with previous research on VGI quality. Chapter 4 presents the methodology of this research. Beginning with the gaps in the literature, which form the research questions, the suggested approach continues by including data matching and VGI quality assessment, presented in details. Chapters 5, 6, and 7 present the three case studies, following the same chapter structure. Chapter 8 provides some further validation of the method, discussion and general findings, obtained from all case studies. Implications for VGI and limitations of the suggested framework are discussed, and the chapter ends by linking back to the potential usage of this research, which is already started as discussion in sections 1.5 and 1.8. Chapter 9 concludes the research, linking the results with the research aim and objectives and suggesting future work and further research.

Finally, Appendix A provides a description of the application developed in PHP to perform the quality analysis, using the necessary screenshots. Appendix B complements VGI literature review by including additional VGI characteristics that are not directly related to data quality. Appendix C provides further examples of data completeness evaluation from all case studies.

Chapter 2: Literature Review

Volunteered Geographic Information (VGI)

2. Volunteered Geographic Information (VGI)

2.1. Introduction

This chapter aims to provide a detailed analysis of Volunteered Geographic Information (VGI), focusing on its aspects that are related to the research area of this study. Starting from defining the phenomenon, examples of VGI projects are presented and VGI characteristics are analysed, especially those affecting VGI quality. Finally, quality issues of VGI raw data are presented.

2.2. The emergence and definitions of VGI

For many centuries the production, dissemination and sometimes even use of geospatial data was considered to be the expertise and profession that demanded specific knowledge, which few people had. Nowadays this seems to be changing. Map makers struggle between:

a. the evolution of technology that provides new surveying and map making tools (which although making their work easier, they may require training and a radical change in a chain of production that until now worked fine),
b. rapid infrastructure growth (including road networks, city expansion, other human constructions) that makes maps obsolete sooner than before or, even worse, sooner than the planned frequency of updates,
c. the increasing demand for new types of data for novel applications that represent forms of information that was previously uncharted (ranging from maps close to their traditional meaning and use, such as cycling maps or city maps for disabled people, to applications monitoring social phenomena or natural disasters),
d. the difficulty in answering this demand by producing all these new products and at the same time retain the quality standards on which their reputation, validity and high quality relies,
e. the decreasing funding for the production of maps (Goodchild, 2007a), whether the provider is an NMA and depends on government funds or a commercial provider that depends on sales and market demand. The cost of spatial production, on the other hand, may increase despite the technological improvements of the instruments: although they make data gathering much easier and faster, more detailed information than before is desired. In most cases this information has a local sense and demands visiting the place to collect it. For example, a satellite image covers a large area and can be digitised in order to create a road map, but offers no information about the road names. Thus, the map provider has to consider that apart from the original cost of the image, the cost of some field work to

georeference the image and the cost of digitising, some additional field work is needed to gather local information such as road names, points of interest, etc.

Users, on the other hand, have always been in need of spatial data specifically fit for their use, which usually differs from what is commercially provided. Until recently they were forced simply to compromise with what they had been given, usually not without paying a considerable price. Recently, however, they evolved and changed their attitude as a result of many reasons:

a. Compared to two decades ago, an increased education level (people reach higher academic levels) and a technologically advanced way of life, allows someone to conceptualise simple geographical terms and to be able to read (or even compile if given the necessary tools) a simple map, whereas in the past a degree of expertise was necessary. According to Goodchild (2008a, p.2), '*Everyone feels himself or herself to be an expert in geography because geography is experienced by everyone*'. He adds that although there are certain areas of the planet and types of geographic information that require advanced skills and thus they can only be addressed by professional cartographers, there is also simpler GI, which is now possible to be produced by almost anyone; '*Mapping of streets and other well-defined features may require simple skills that almost anyone possesses: the ability to use GPS to determine location, and the ability to identify the names and other obvious characteristics of features*' (p.6).
b. The technological advancements in simple positioning devices (such as GPS), helped by the removal of the deliberate GPS's signal degradation (called Selective Availability) in 2000 by US President Bill Clinton, allowed people to be equipped with positioning devices of accuracy of a couple of meters that meets the average user's needs. Nowadays, with GPS technology implemented in most mobile phones, one can be provided with a positioning device even if this was never among one's consumption priorities.
c. Web 2.0 technologies (such as Application Programming Interfaces (APIs), the use of Asynchronous Java Script and XML (AJAX), the client-server architecture) and the evolution of desktop GIS to internet GIS, allow someone to interactively use different mapping software programs, and to combine them in order to produce a customizable spatial product or software for himself. Sui (2008) describes it as 'wikification' of GIS. Although certain skills are required in order to achieve a 'mashup', as it is called (explained later on), it is not necessary to program or even understand the language of the combined software. The first example of such a user evolution back in 2005 is Paul Rademacher's Housingmaps.com, which combines information of a website with apartments and houses for sale and Google

map (Housingmaps.com, 2012). Other user mash-ups can be found in Google Maps Mania (2010) website.

d. The increased level of geographical understanding in combination with the increased demand for a specific type of updated spatial data for certain needs may drive users away from the standardised and typical spatial products provided by NMAs or commercial MOs, especially in cases where the price is high. Average users with no GI expertise have little or no understanding of quality standards. The price for a spatial dataset which does not always contain all the information they want, or contains additional information not important to them, or is not recently updated, will always be considered as unjustifiably high, regardless of its quality standards and existing metadata.
e. The detailed information needed for some applications usually refers to indigenous experience and is not mapped by any NMA or commercial providers that compile and produce spatial datasets for a distant place without always visiting it (e.g. routing applications, finding addresses, etc).

The result is a new trend in Geography, with more social than technical characteristics, which was born out of the above mentioned factors rather than scientifically discovered and developed as a result of specific technological advancements. Certain aspects of Geography and Cartography seem to pass from the hands of experts to the hands of simple users, turning them from users to producers, also named 'produsers' by Bruns (2008). Researchers have tried to define the new trend and / or its derivatives using various terms, mentioned in section 1.1. In order to be able to conceive what this trend is, a selection of the above definitions will be presented.

According to Egenhofer (1995), **Naïve Geography** is the common-sense geographic knowledge that an average citizen with no GIS training has, regarding a relatively limited surrounding world in which he or she constantly moves around. This knowledge is used almost instinctively to solve everyday tasks, such as which road could be a shortcut for a certain direction or a customised orientation based on landmarks. He further uses negation to better describe the meaning of naïve geography; it is not geography by or for the illiterate, stupid or simple-minded (p.5). Qualitative instead of analytical methods are often used and naïve geographic reasoning can be inconsistent. He relates it with other disciplines and he argues on some of its basic elements. Although Egenhofer (1995) points out the need to model Naïve Geography in a GIS environment, this knowledge will remain strictly instinctive and personal. Recently, the above mentioned technological advancements have provided tools to express part of this personal geographic reasoning, e.g. someone could create a

map of his surrounding area in the way he or she conceives it; however, despite that the personal conception is thus expressed to a wider audience, its adoption remains uncertain. Although 'Naïve Geography' is not likely to be a suitable term to fully describe a trend which would appear a decade later, it could be considered as the beginning of an effort to define part of it mainly by its argument that there are some basic spatial skills in everyone.

Turner (2006, p.1-2) gives a definition of **Neogeography**. '*Neogeography is about people using and creating their own maps, on their own terms and by combining elements of an existing toolset*.' He distinguishes it from the traditional ways professional cartographers follow; the software, the output formats, the troubling questions are far more different, as well as the purpose of a map. For example, map projection is unlikely to be a matter of discussion between neogeographers, it is even possible that they will have never heard of the term. On the other side, fun is unlikely to be the reason to create a map for professionals, and terms such as 'geotag', 'mashup', are also possible to be unknown from their side. Currently, however, these questions may partially be outdated: the above terms have become familiar to each side.

A derivative of Turner's definition is the term '**mashup'**, with origins in the music industry. A web mashup is a combination of two or more web pages, online data or web services, in order to provide a new one, by using provided APIs. An example is Flash Earth (2010), a mashup created by combining APIs provided by Google Maps, Microsoft Virtual Earth, NASA World Wind, OpenLayers and Yahoo! Maps, according to which a user can zoom in an area and switch the view between the above maps. According to programmableweb (2010), 46% of all mashups are geospatial (called 'mapping mashups' by programmableweb). Although mashups are part of the new trend as a result of the recent technological advancements, they do not rely on the contribution of a number of users; they combine existing data and they are created by one or few persons with the necessary skills.

Goodchild (2007a) uses Estes and Mooneyhan's 'mapping myth' to show that the world is not well mapped; maps are not constantly being updated or made more accurate. He states that the declining government funding for mapping purposes in many countries (including the U.S.), can be dealt with a new means of acquiring geographic information, which he aptly names '**Volunteered Geographic Information**' (VGI). His view of a world comprised of six billion human sensors, that can provide unique spatial information for their local environment to supplement the traditional mapping tasks of NMAs, may seem too optimistic or theoretical to be achieved at a worldwide level, yet it is not impossible for smaller areas. Although he mentions that VGI and traditional mapping are

very different, a fruitful combination of these two worlds sounds interesting, although he also gives a hint of some implications, such as the difference in existing structures or the flow of information and data quality. A lot of research is needed towards that direction to deal with these implications as well as with others, such as copyright issues. However, his definition appears to be suitable to describe the new trend's nature of apparently pure and selfless contribution of spatial data by individuals.

Howe (2006) uses an example in his article to show the meaning of '**crowdsourcing**', a word coined by him. Although not specifically referring to GI, he shows how a specific task can be carried out by a crowd through an open call for contributions, significantly reducing the cost if compared with assigning the same task to one or more employees. However, his definition follows a different path from Goodchild's term of VGI, not only because it does not refer specifically to GI, but also because it implies an exploitation of individuals' contribution from a commercial or business body for purposes of making profit.

Another term which seems to be related to VGI and crowdsourcing, is **Public Participatory GIS** (PPGIS), which emerged in 1996 during the meetings of the National Center for Geographic Information and Analysis (NCGIA). Realising that GIS can provide tools that can lead to exploitation and marginalisation of public communities, PPGIS aims to use GIS to empower these excluded communities through access to spatial data, education and participation, so that they can have their own voice in decision-making processes that concern them. Participants should be involved in the creation and evaluation of data. Different values and views that lead to contradiction and inconsistencies are welcome, since they can prevent a premature decision, and the final output should reflect the participant's goals. However, when more than one communities are involved, some choices may result in disempowerment and marginalisation of one group, and keeping the balance is a difficult challenge for PPGIS (Onsrud and Craglia, 2003). Examples of PPGIS in action can be found on the CRSSA (2012) website of Rutgers University. PPGIS is different from Howe's (2006) 'crowdsourcing' because it does not have a commercial orientation. Flanagin and Metzger (2008) link PPGIS with VGI by offering a view of VGI as an extension of PPGIS. Sui (2008, p.4) also views VGI as PPGIS 'with a much enlarger public', yet he admits that there is a need for research regarding potential implications in privacy and democracy of PPGIS. On the other hand, Tulloch (2008) argues that although there are some similarities or rather blurry boundaries in some cases, there are also differences. In comparison with VGI (or at least the ideal form of VGI), PPGIS usually serves a subset of the public; people participating and contributing do not always have access to the final product;

since PPGIS is mainly directed or organised by decision-makers, among those who benefit may be organisations or government agencies and not simply everyone. Sieber (2007, p.2) states that 'participation' assumes an official process. In other words, PPGIS includes data contributed by a group of people after the contribution is asked, framed and guided towards a goal in favour of the local community by an organisation. Compared to VGI or crowdsourcing, which mostly refer to data collection, PPGIS is a larger concept that also includes data management, analysis and visualisation. However, although PPGIS is not VGI, VGI could act in a similar way as PPGIS in cases of natural disasters (Goodchild 2007a; Haklay 2010b; Mullins, 2010).

Antoniou *et al.* (2010b) use the term '**User Generated Spatial Content**' (USGC), focusing on the spatial meaning but also extending to a more general level, so that they can include volunteered, as well as other crowd-sourced methods of collection. Although suitable, this definition is not yet as widespread as the term 'VGI'.

For this research, the most appropriate term seems to be Goodchild's (2007a) 'Volunteered Geographic Information' (VGI). Using OSM project as an example (described in the following section), people indeed volunteer to provide geographic information to the project, which in return can be accessed by anyone (regardless of one's previous contribution). The information provided can also be characterised as local in many cases; street names or points of interest such as pubs and restaurants are only gathered by inhabitants or visitors, who respectively have indigenous experience or gain local knowledge and can be considered as human sensors. The 'neogeography' term, although also applicable and relevant in a sense that the OSM's contributors are 'average citizens' (Egenhofer, 1995) instead of 'professional cartographers' (Turner, 2006), can be considered a more general definition. 'Naïve geography' refers to the basic and instinctive geographic skills everyone has, so it is not a suitable term. 'PPGIS' is not suitable either because of its different goals. 'Crowdsourcing', finally, can be partially accepted, in terms of the way data are gathered.

2.3. Some examples of crowd-sourced or VGI projects

A broad category of crowd-sourced information is the one referring to photo sharing web sites. When contributed photos are geo-tagged, in other words contain coordinates as part of their metadata, these projects can be considered as VGI. Such websites include Flickr (originally known as Yahoo! Photos), Picasa Web, Panoramio, Geograph, Webshots, SmugMug, TwitPic. Antoniou *et al.* (2010b) extensively examine the spatial aspect of the first four. For this category, though, the GI provided by volunteers is restricted to geo-tagging. Other websites, however, demand and offer

richer GI. An example is Everytrail, a web site that allows users to connect their uploaded photos with GPS tracks, add a story and share it through the web. Moving to more complex GI, another broad category of crowd-sourced projects with explicit geographical content refers to map editors. Some of these projects are:

Google Map Maker is a proprietary project, based on crowdsourcing. Data input is limited to digitization of Google satellite imagery. Users digitise the images and add or edit features (such as roads, Points of Interest (also called POIs: restaurants, banks, hotels, etc), polygons), populating a global, spatial, vector database. Creating a user account is necessary for editing, yet editing is feasible only for 190 countries (Google Map Maker, 2012). Descriptive attributes need to be selected from a domain. This implies a range of values to cover most cases, so the user is somehow limited in describing the information added. However, this leads to a more structured database. Contributions need to be approved before appearing online. According to the terms of Service, the volunteer has no ownership over the contributed data and there are some restrictions regarding the data usage (Google Map Maker, 2011). Personal or others' contributed data can be viewed, and recently a policy change also allows downloads. Data export format is limited to shapefiles. As it is a proprietary project, data download is only possible in selected countries (Figure 2.1).

Figure 2.1: *Google Map Maker data availability (Source: Google Map Maker, 2010)*

Map Share is a crowdsourced spatial database, also proprietary, which was started by TomTom, a navigation systems manufacturer. Those belonging to the TomTom mapping community can make instant corrections to their map in their TomTom device, and by occasionally connecting to the TomTom Home webpage, they provide these updates to the TomTom mapping community, while at the same time they receive updates from other users. Editing through digitisation of a satellite or aerial image is not possible, so updates are made in the field. There are options for the use of updates, such as to use personal changes, and/or changes verified by TomTom, trusted sources, reported by many or some, etc (TomTom Map Share, 2012). Users can also pay to advertise themselves (Yourtomtom, 2010). The TomTom mapping community is not open to anyone; one needs to purchase a TomTom navigation device, as well as a Map Share compatible map. The created map remains the property of TomTom and cannot be downloaded in other formats or used otherwise. However, TomTom is launching OpenLR encoding technology, described as an open, compact and royalty-free dynamic Location Referencing, which will provide a new interoperable map format (suitable for other devices as well). One of the disadvantages of location referencing is that it needs identical maps at both sides of the communication (TomTom, 2010), otherwise there may be inaccuracies. This might not necessarily be a disadvantage for TomTom, since it adds value to their maps and makes them irreplaceable.

Wikimapia is an editable, interactive, global map which uses Google satellite imagery as background. As its name implies, it is not proprietary; it is a user created project, following the wiki style, which aims to '*create and maintain a free, complete, multilingual, up-to-date map of the whole world*' (Wikimapia, 2010a). According to its terms of service and in comparison with the previous examples, the content voluntarily uploaded by users is made public and can be used by everyone, however only for non-commercial and non-public reasons, otherwise a special agreement is required (wikimapia, 2010b). As with the previous examples, there is no option for downloading data, however this is not due to proprietary reasons, but because of different objectives of the project.

People's map is a similar project, free from third party copyright, which allows users to create an individual map of Britain, using online tools to digitise Getmapping's aerial photographs. Membership activation is needed, during which the user is informed that his contribution will belong to the People's Map project. Data are free for private and non-commercial use, while for professional use they are licensed in perpetuity (People's Map, 2010). Contributed data by average users or professionals can be viewed by all those who access the website. There is a procedure of data verification, which takes time, so contributed data are at first characterised as 'not verified'.

The user has the choice of viewing all the data or only the verified one. After data are verified by People's Map Partnership, it is integrated into high quality digital maps. Although the final product is available in different formats (raster image, vector layers, shapefile), users have to pay for it. Anyone who wants to create their own maps can buy the base map from People's Map and customise it afterwards by adding information or changing the style through a provided API. So, this project gives access to the contributed data, however not for free, and only for Great Britain. Recently the project 'went offline' (People's Map, 2012), but the previously created products are still available.

OpenStreetMap is a global map editor based on user contribution regarding uploading of GPS tracks or digitisation of satellite images provided by Bing and others. The main difference with the previous projects lies in the ability to freely download the contributed data in various formats, regardless of being a registered contributor; the contributed data belongs to the anonymous users, who do not have to pay to participate (like MapShare) or to download data (like People's Map) (OpenStreetMap, 2012c). Based on the same general context of crowdsourced GI, the fact that data are free to download makes it ideal for research on VGI aspects, because some conclusions could be generalised to cover other VGI projects. This is the reason to use it in this thesis, hence more details for the project are provided in the next section.

2.3.1. VGI Source: OpenStreetMap (OSM)

Since OSM is used as a VGI source in all case studies of this thesis, a more detailed description follows to allow for a better understanding of some technical details and to justify the reasoning behind the way the comparison methodology was developed.

History of OSM

OSM is an open source VGI project that '*creates and provides free geographic data such as street maps to anyone who wants them.*' (OpenStreetMap, 2010e). OSM's characteristics of a main server to hold the data, tools to edit them, a network through which editing is possible and a number of dedicated contributors who act as geographic sensors, form a clear example of Goodchild's (2007b) notion of VGI (Ather, 2009). OSM started in the UK in August 2004 by Steve Coast, as a result of his frustration with the difficulties users face when they need to acquire, process and further distribute data from the OS, GB's NMA (Chilton, 2009a). However, the project's area of coverage is extending rapidly every year (Chilton, 2009a) and soon exceeded the UK borders. This on one hand is the result of the increasing number of volunteers (OpenStreetMap, 2010a), and on the other hand the result of massive data import, such as USA TIGER Data, the entire street map dataset for the Netherlands and

the road networks for China and India (Chilton, 2009b), donated by relevant providers. The significance of the project in the VGI area can be seen in Chilton (2009b), who mentions the future imports that are already under way in order to be integrated, and by the increasing number of web mashups that use OSM data as one of their components. An example is PhtoSM (2010), which combines Flickr photos and the OSM dataset in the area of Haiti to support aid providers after the earthquake. Switch Maps (2010) is another mashup example with which you can switch the background map between Google Maps, Google Earth, Street View, Bing Maps, OSM and Yahoo Maps (without having to open new windows and manually reselect your area). New user demands inspire developers to extend the use of OSM by inventing new applications (a list is available in OpenStreetMap, 2010j).

Data coverage is not complete and, although it is growing fast, rural areas still remain scarcely mapped (Haklay, 2010c, p.11). However, the purpose of the project is not to cover the whole world (Haklay and Weber, 2008). Considering the dynamic aspect of VGI, OSM will never finish as long as there are contributions altering the data. On the other hand, Chilton (2009a, p.4) shows the unrivalled up-to-date data of OSM in some areas, with an example of Heathrow airport in London; on the day the new Terminal 5 was open to the public, OSM already had related road information available.

An additional aspect of OSM as a VGI project, rather hard to find in other crowd sourcing projects, is the ability to retrieve data for free regardless whether someone contributed to it or not. This makes it the most valuable source to study VGI. Following Haklay's (2010a) classification, it is a non-profit egalitarian VGI project.

How OSM works

OSM web page contains 3 main parts (Figure 2.2). On the upper side of the map frame there are 6 tabs, which allow the user to edit data ('Edit'), import data ('GPS Traces'), Export data ('Export'), or view history of changes, user diaries or the map (tabs 'History', 'User Diaries' and 'View' respectively) (OpenStreetMap, 2010e). Help and Wiki pages contain a lot of information and are very helpful for new users, including also videos on how to use the main functions of OSM.

Haklay (2010d) provides an analytical presentation of the OSM project, referring to its background, the editing tools provided, the technical infrastructure, the social collaboration through mapping parties and its challenges. Singer (2009) moves deeper into more technical details on how OSM

works with postgreSQL. A more detailed description of OSM can be found at Ramm *et al.* (2011), who target all possible audience, from beginners in mapping to web developers.

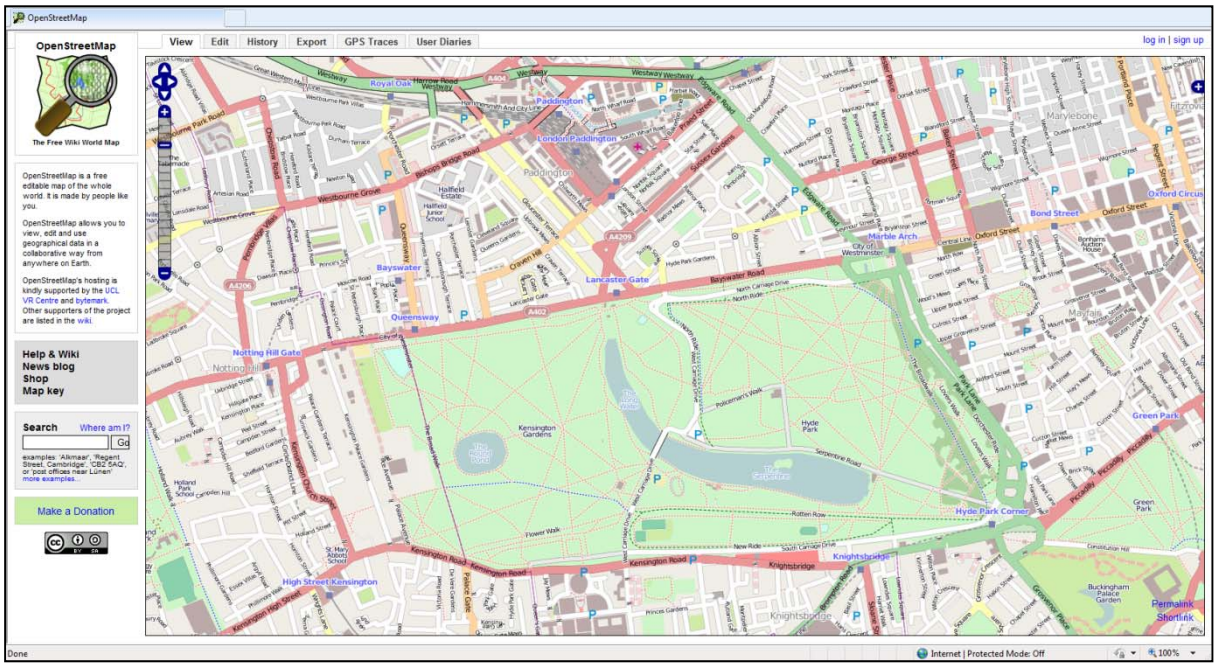

Figure 2.2: *OpenStreetMap webpage (www.openstreetmap.org)*

The 'elements' or entities used by OSM to represent reality are 'nodes' (meaning points to describe POIs or road junctions), 'ways' (meaning lines to describe road segments) and 'relations' (meaning groups of objects with a specific role, for example restrictions, boundaries, multipolygons). A polygon area can be represented by a closed way. Although there are suggestions on how to add descriptive attributes, called tags, users are free to choose their own tagging methods (OpenStreetMap, 2010d). However, as Sieber (2007, p.1) mentions, '*OSM founders are quite transparent about the messy condition of their metadata*', obviously referring to the OSM statement that '*There are no real standards in OSM, the only thing that is defined is how to get data from OSM. That data can be created in many different ways, in a hope that entropy will fix things with time.*' (OpenStreetMap, 2010f). Inevitably there is a trade-off between the volume of data adequately described by metadata and data contributed by volunteers. As a result, OSM data are heterogeneous in all quality aspects, such as data completeness, positional accuracy, attribute completeness and consistency.

There are four ways to contribute to OSM. One is to go outside with a GPS device, move around and upload the GPS tracklog (in GPX format), preferably with some relevant information (e.g. street names). People can do this individually or can form groups to map a specific area, usually called

mapping parties, like the one that mapped central Manchester in one weekend (Perkins and Dodge, 2008). A second way is to digitise a raster image; the user can select between global satellite imagery provided by Bing Maps since 2006 for the purposes of tracing (OpenStreetMap, 2010g), or existing maps for certain areas, such as StreetView for the UK (released by OS in April 2010) or NearMap for Australia. A third way is to import a large amount of data, usually provided by a MO or NMA as a donation. This can be achieved after some strenuous but necessary processing in order to integrate the new data in the OSM database, usually carried out by the core programmers of OSM (which is also a group of volunteers) (Chilton, 2009b, Haklay & Weber, 2008). Finally, the fourth way is through paper map. Walking-papers (2010), a web application developed in 2008, permits the user to download and print a map, move around and draw on it what is not mapped, scan it and upload it updated (Walking-papers, 2010) so that someone else will digitise it.

Data adding is feasible through a variety of editors, the most important of which are Potlatch and JOSM. Potlatch is an online Flash-based editor, which is relatively easy to use and addresses the general user. JOSM (Java Open Street Map) is a heavier editor, which allows offline editing and provides more functionalities than Potlatch, permitting the import of large chunks of data. Other editors are Merkaartor (stand-alone cross-platform editor), OSM2Go (for mobile devices), Vespucci (for android devices), Amenity Editor (for nodes and Points of Interest, known as POIs), MapZen (a flash-based editor created by CloudMade) and other editors for iPhones (OpenStreetMap, 2010h; Ramm *et al.*, 2011).

A user registration is necessary for adding or editing data. Data are stored in a database designed to support a wiki-style behaviour, which means that versioning is enabled, rollbacks to previous versions are possible and no previous information is deleted. Access is possible through a RESTful API, allowing mashups and development of tools independently from the database. The servers containing the database are hosted in the University College London (Haklay and Weber, 2008).

The important aspect of contributed data is that it must not be copyrighted, but gathered either by the users' own effort or by other free source of data. OpenStreetMap (2010b) states in bold letters '*Do not use data from copyrighted maps or any other proprietary data!*', explaining why, and ending the paragraph with the sentence '*If unsure: do not use*'. Users cannot add street names from a map they bought, or digitise an aerial photograph they purchased from a relevant provider, or upload vector data sold by other map providers, otherwise copyright issues will arise and legal claims will be placed against the OSM project. OpenStreetMap (2010c) so far uses 'Creative Commons Attribution-

ShareAlike 2.0' (CC-by-SA), mentioning that '*If you want to use OpenStreetMap, you will have to give credit to both OpenStreetMap and the license*'. This type of license is what McConchie (2008) refers to as 'copyleft' (to oppose it from the term 'copyright'), which allows the remix and reuse of the data provided, but also demands that the derivatives will also be free to be remixed or reused in the future, preventing someone from claiming ownership of a spatial product based on free data. Although present licensing allows the derivatives of OSM data to be shared (Singer, 2009), it can also prevent the use of data. As a result, after more than three years of discussions, OSM moves to a new copyright framework (OpenStreetMap, 2012b) in the direction of adopting an Open Database License (ODbL) (Chilton, 2009b).

As output, there is a growing number of available formats among which one could choose to download data from OSM; XML, raster image (PDF, JPG, PNG format), vector data (SVG, KML, SHP format), etc. There are instructions available on how to download data in less common formats (e.g. for Garmin GPS devices, Manifold GIS software, ESRI's shapefiles, etc) or how to download data for relatively large areas (e.g. in national level) (OpenStreetMap, 2010i). There is no need to register in order to export data (in comparison to when importing data).

Although OGC provides a standardisation method that can explicitly deal with interoperability as well as implicitly address heterogeneity through open GIS standards, OSM refuses to comply with OGC standards for reasons of simplicity, maintenance and lack of supporting the wiki-style behaviour of the project (Haklay and Weber, 2008). Although there is no top-down quality assurance process for OSM (Haklay *et al.*, 2010), there is a growing list of 'Quality Assurance' tools (OpenStreetMap 2010k) that provide a list of bugs in the data. These can be 'hopefully' manually fixed by users' editing afterwards. Such tools are OpenStreetBugs, Keep Right, Osmose, Maplint, Way Check, MotorwayCheck, Duplicate Nodes and many more. Most of them address digitisation errors, such as ways without nodes, non-closed areas, open ends, missing tags, almost junctions (called 'undershoots' in cartography, meaning lines that do not intersect although they are supposed to), unmapped places, etc. However, these tools mainly point out gross point positional errors or lack of attributes; they cannot address linear positional accuracy, they cannot provide quantitative information on the quality of OSM data and they rely on the user's will to fix the errors.

The need for quality evaluation of OSM is mentioned by Goodchild (2008a, p.10), who argues that '*one way to establish authority would be for novel sources such as OpenStreetMap to (…) initiate programs of quality testing*'. Haklay & Weber (2008, p.17) also mention the same research problem:

'*some idea about information quality is crucial for evaluating how fit OSM data is for various applications*'.

2.4. VGI Characteristics

As VGI started to expand in its various forms and applications, it has attracted the attention of many researchers, who apart from simply defining it, moved further to discover its objective, its implications, as well as the barriers and how they can influence its prospects. In general, researchers who addressed the VGI phenomenon can be divided in two categories; the supporters and the critics. In an effort to grasp the meaning of VGI, the following section will review their thoughts and arguments by commenting on perspectives, data quality, quality standards, metadata, heterogeneity, credibility and other issues regarding VGI.

2.4.1. Perspectives of VGI

Various VGI projects started for different reasons and, as time goes by, their objectives either remain the same, or transform to adapt to new users' demands. Their uses can extend the ones they were designed for. As an example, Flickr is a website that offers people publicity (by enabling them to freely publish their photos), as well as new ways of organising photos and video (Flickr, 2010a). Yet, geotagged photos are a growing database that can offer worthy spatial information (Flanagin and Metzger 2008; Antoniou *et al.*, 2010b). Other projects, although relying on contributors having fun, they have more specific motives and objectives. For example, OSM provides digital spatial data without the technical restrictions on their use that could hinder creativity and productivity (OpenStreetMap, 2010e).

The flexibility and prospect of creative combination that Web 2.0 offers can also lead to new perspectives and uses of VGI that were not originally anticipated. Additionally, apart from the user's personal interests and reasons for contribution, VGI can also be directed to larger communities in need of help. An example provided by Chilton (2009a) is the case of Israeli and Palestinian conflict in Gaza and the lack of up-to-date data, which was efficiently supplemented by the VGI project of OSM. Another example was given in DGI (2010) conference by Jeff Peters, Director of ESRI Federal programs, who mentioned during his speech that when Haiti earthquake occurred, map mashups and VGI helped people to notify rescuers about trapped people while official maps were not available, noting also that OSM proved to have the best available vector data for the area. Mullins (2010) examined crowd-sourcing in Haiti and also reaches the same conclusion, adding that since

Katrina disaster, crowd-sourcing has grown far more effective through centralisation. A more detailed description of the Haiti response of OSM is presented by Haklay (2010b), who compared VGI provided by OSM with that of Google Map Maker. The use of VGI for early warning in case of natural disasters is also mentioned by Goodchild (2007a).

The use of crowd-sourced spatial data by decision-makers is noted by Harrison and Haklay (2002) and Seeger (2008), who respectively comment on their use in the environmental and landscape planning and site design process, providing thoughts on how decision-makers could use public participation. Following McLuhan's Law of the Media, Sui (2007) examines the areas of social practise that VGI can enhance, make obsolete or revive. From the perspective of Usability Engineering, Haklay (2007a) uses Roger's model of diffusion of innovations to focus on the way users adopt a new application. Following more technical approaches in the direction of integrating VGI, researchers such as Craglia (2007), Budhathoki *et al.* (2008), McDougall (2009), argue about ways to combine VGI architecture (or lack of it) and existing Spatial Data Infrastructure (SDI).

The dynamic aspect of VGI in terms of users' contributing whenever they can, however they can (being at home or visiting the area) and at marginal cost, leads to the creation of an up-to-date spatial database, which potentially will always be updated as long as users contribute. Conventional spatial databases do not have this advantage; there are funding limits that frame and restrict the update policy, and time latency is likely to be present between the gathering of data, the processing and the final output. For example, OS uses the government funding to maintain an update policy of inclusion of significant changes within six months (Edina, 2012). However, only random or specific areas are checked all over the UK and are being updated every six months, not the whole country (Coote and Rackham, 2008). Goodchild (2007a) views humans as sensors that possess local spatial knowledge and suggests – at a theoretical level – that VGI can be used to update official maps that no longer can be updated by NMAs, due to high costs and lack of local information that cannot be extracted from a satellite image. Integration of VGI in an existing spatial database is an important perspective, although Goodchild's theory (2007a) cannot be easily put into practice without taking into consideration VGI implications (such as copyright issues or heterogeneity), discussed later on.

Among researchers in favour of VGI, Goodchild (2007a) predicts that VGI will eventually replace traditional mapping as a centralised process. Goodchild (2008b, p.242) adds that '*all three arguments for mapping expertise—the need for cartographic skills, skills in the operation of complex measuring instruments, and familiarity with the subject matter of mapping—may have disappeared*

as a result of technological change, and not only for conceptual simple types of geographic information such as placenames'. However, some others have their objections (Walsh 2008; Budhathoki *et al.*, 2008) due to lack of data quality standards and of expertise irreplaceable in some aspects of professional cartography. Thus, the view that VGI will generally replace traditional mapping should not be accepted; even after developing adequate quality assessment mechanisms for VGI, it still serves different purposes and address different spatial needs than traditional mapping; yet, under certain conditions they may complement each other.

Due to the variety of VGI perspectives, the need to understand it, evaluate it and use it or integrate it in a wide range of applications is essential, taking into consideration and assessing the criticism already present in literature. Evaluation stands in the middle of the procedure and is an important step, before the 'fitness-for-use' examination of VGI or its integration with existing databases.

2.4.2. Data quality of VGI

The vast Geographic Information gathered by volunteers urged researchers to tackle the problem of data quality of VGI in both its forms; the low quality (either assumed or proven) and the lack of tools and methods to estimate it. This shows that VGI can no longer be considered something ephemeral or without consequences in GIS. The need for assessment of data quality in GIS (subjectively or quantitatively) emerged in the past after people started using GIS and realised its power and perspectives, as it will be explained in the next chapter. In a similar way, data quality knowledge now becomes a necessity for VGI to those who want also to use these data in science or decision-making, apart from personal and trivial purposes such as planning their vacation. However, until now there has not been any systematic analysis on the quality of VGI (Haklay, 2010c, p.1), and the fact that its quality is not guaranteed hinders its use despite how 'attractive' as a source of free data is (Mummidi and Krumm, 2008; Brotzman, 2009).

VGI cannot satisfy professional GI needs with inflexible requirements in terms of completeness, timeliness or positional accuracy (Budhathoki *et al.*, 2008). Dodge and Perkins (2008, p.4) call these products 'Mc-Maps', aptly comparing them to fast food. However, this does not apply to all VGI projects; OSM has a relatively high quality (Haklay, 2010c) despite the lack of quality standards, existence of sufficient metadata or an internal quality assessment mechanism.

Data quality is a matter not always conceivable by users, especially the non-experts. As Kraak and Brown (2001, p.45) mention, '*Users are not always fully aware of the quality of the maps; as long as*

the map looks nice and detailed, they often think that it is their fault if they cannot find their way with it easily.' When these users start to contribute to VGI, quality of the provided data can be questionable. As a result, data quality and ways to evaluate it may be among the biggest limitations of VGI. In fact, many of the characteristics or other drawbacks of VGI, that will be mentioned later on, are the source of data quality issues, and by finding a mechanism to evaluate VGI quality, these problems can also be dealt with, or at least be reduced. In other words, a way to assess data quality of VGI can reveal where and which data are erroneous, answering the fit-for-purpose question and negating at the same time problems stemming from lack of metadata, lack of credibility, misuse of VGI regarding the deliberate flow of misinformation, or definition of real motives for contribution. This renders the quest to find answers for all the factors that lead to these problems rather insignificant for a fit-for-purpose examination.

A number of researchers argue on the need to address the data quality issue of VGI (Flanagin and Metzger, 2008; Coote and Rackham, 2008; Auer and Zipf, 2009; Antoniou *et al.*, 2010a). Apart from the theoretical approach of data quality, a lot of researchers argue on the importance of finding assessment methods and tools to estimate VGI quality (Goodchild, 2007a; Sieber, 2007; Flanagin and Metzger, 2008; Coote and Rackham, 2008), while others move on to assess the spatial data quality in specific VGI projects and selected areas (Haklay, 2010c; Zielstra and Zipf, 2010; Girres and Touya, 2010; Ludwig *et al.*, 2010).

In order to determine VGI quality in quantitative terms, a comparison with official data of higher quality (meaning data produced by more accurate instruments or following accepted quality standards) could provide an answer (Goodchild and Hunter, 1997; FGDC 1998a; FGDC 1998b; Devillers and Jeansoulin, 2006; Maué and Schade, 2008). However, this is not always applicable, since VGI may include information that has never been mapped before (Maué and Schade, 2008). Also, official data are simply data of higher quality; although they are usually referred to as 'true' or 'ground truth' data, it does not mean that they represent the real world correctly. '*No geographic data can be perfect, since it is based on measurements and observations and subject to innumerable sources of uncertainty.*' (Goodchild, 2008a, p.8).

Of course, due to the variety of VGI, it will neither be possible for a method or tool to be implemented on all VGI projects and deal with all the aspects of each one, nor to address all the problems already mentioned. However, it could provide a tangible way to express quality in a specific type of VGI in measurable units, allowing decision-makers to decide whether to use the data

or not for their purpose. This thesis is aiming to address VGI quality by proposing a systematic way to assess its spatial quality in reference to a dataset of higher quality. Therefore, a separate chapter will be devoted to data quality literature. This section continues by focusing on other aspects of VGI that influence data quality.

2.4.3. VGI quality standards, metadata and heterogeneity

The need and the role of metadata in GIS are addressed by many researchers. However, although metadata are partially the result of a need to assess data quality, research shows that they are not applied as they should. According to Agumya and Hunter (2002), although metadata standards exist, individuals do not know how to use it in order to see if the data are fit-for-purpose. Goodchild (2008a) mentions that individuals do not pay attention to the published metadata if and when they exist, and in other cases they '*automatically ignore it*' as confusing (Boin and Hunter, 2007, p.6). Other researchers note the reasons that although metadata standards do exist, they are not applied (Tveite and Langas 1999).

VGI applications in general have a loose policy on metadata, which is somehow inevitable; the desired simplicity and lack of users' expertise lead to the omission of metadata referring to data quality. As Sieber (2007, p.1) mentions, '*updating the information or providing the metadata tends to be cumbersome and less glamorous than the initial release of a product*'; forcing the users to upload their data in a specific way will drive them away, since it will be far more difficult and less interesting for them. As an example, the result of a comparison of four photo-sharing web sites (Antoniou *et al.*, 2010b), shows that the web site with some limitations posed on the data collection and structure attracts less people.

The lack of metadata and quality standards is attributed to inability due to the sheer volume of information and lack of incentive, as opposed to a government agency that has the authority and the obligation to do it in order to maintain its credibility and reputation (Flanagin and Metzger 2008; Goodchild, 2007a). Some groups of users or simple data consumers do need a quality statement by VGI projects (Coote and Rackham, 2008; Boin and Hunter, 2007). However, Flanagin and Metzger (2008) admit that this is something a user should be able to decide based on existing metadata or other quality standards.

Usually the metadata provided in VGI are a combination of the most significant (according to the designer) implicit metadata and of some or no explicit metadata. Whereas implicit metadata (such

as the username or time of contribution) can be created automatically, provided that the application is designed accordingly, explicit metadata (such as the nominal accuracy of the GPS device used) relies on user's manual editing. Poore and Wolf (2010) compared the metadata of professionals with that of VGI, showing that the second are explicit-dominated, in comparison to the first. However, although on one hand implicit metadata reduces the amount of work, it is more trustworthy and allows for a more cohesive interaction between systems, on the other hand it demands knowledge of its existence, it lacks flexibility and it does not help communication between the user and the developer (Ehreke, 2006), which are necessary aspects in VGI.

The selection of implicit metadata chosen to be presented and the avoidance of others may rely on reasons such as metadata heterogeneity of the data subsets, commercial policies of classified information or poor selection due to an inexperienced designer. Google Map Maker, as an example, which uses Google Earth satellite imagery, does not provide information on the spatial accuracy of the information included in its various layers (Goodchild, 2007a). The selection of explicit metadata, on the other hand, relies mainly on the users and the desired simplicity. Even if a VGI project is designed to accept the necessary explicit metadata, it is uncertain if the user will bother to fill in the information. An example of '*no real standards*' is OpenStreetMap (2010f), where there is neither metadata about the accuracy of the provided data, nor internal quality assurance procedures (Haklay and Weber, 2008).

Considering that VGI is flowing from different individuals with no or loose coordination or submission to some standards, heterogeneity is expected to be 'tremendous' (Elwood, 2009), not only between various VGI projects, but even within a single one. Heterogeneity is mainly the reason of personal conceptions of how and what needs to be mapped, varying methods of digitisation, different positioning devices and ways to use them, different aspects of data quality and varying levels of effort to tag data or add metadata. Although the Open Geospatial Consortium (OGC) intends to deal with heterogeneity implicitly (by offering interoperability solutions through open GIS standards), following OGC Standards may prove to be difficult or undesired for many VGI projects regarding data creation and metadata, as in the case of OSM (Haklay and Weber, 2008). In other VGI cases this results in inventing their own standards, as in the case of Google Earth API and the KML format. So, despite that research has already started on how to apply open standards to VGI, enhancing its quality and credibility and at the same time not restricting the contribution of volunteers (Auer and Zipf, 2009), an individual contributes more willingly if there are no barriers to

the type and way the information is added; in cases of restricting data structures, the time and effort needed to get acquainted with the existing framework could potentially drive the user away.

With no or insufficient quality standards, metadata, and a frame to reduce heterogeneity, data accuracy and quality is gravely affected, since once a specific VGI contribution is integrated to a larger dataset, data individuality is lost. Everything a contribution can offer, from detailed or accurate to incomplete or erroneous data, is absorbed and unified in a larger dataset of varying and unknown quality. This results in an inability to uniformly describe VGI in terms of quality, in the same way that a single temperature value cannot be regarded as representative of the weather for a whole country. This problem has already been raised in the data quality literature (Devillers *et al.*, 2005).

2.4.4. Credibility of VGI

The main criticism on VGI, affecting its usage, comes from its unknown credibility, which started to be a matter of discussion only recently (Coote and Rackham, 2008; Devillers *et al.*, 2010). Cases of misuse of VGI (such as deliberate misinformation) can additionally affect its credibility. The question is how can we trust data coming from:

- anonymous contributors (with no authority and who obviously cannot be held responsible for possible mistakes), or from 'amateurs working for nothing or for cheap' (Walsh, 2008, p.29).
- people with unknown intellectual background, usually non-professionals in geography or cartography (Tulloch, 2008), who '*in most cases are not trained or even necessarily interested in geography as a science*' (Flanagin and Metzger, 2008, p.139), so unaware of data quality issues or data accuracy theories.
- people who create spatial information with unknown incentives, and the data accuracy or completeness that fits them (according to which they gathered the data) may not be sufficient enough for others' requirements.
- people who use diverse methods and instruments for data collection, hence with different accuracy.
- people or projects that do not include metadata to describe it (Goodchild, 2008a). An interesting example of how lack of metadata can lead to wrong assumptions is given by Keogh and Fraser (2008): many users believe that Google Earth provides real time satellite images.

- projects that mix data from various sources and in the end there is no indication of their spatial accuracy and time of collection or, even further, the original source of data is lost and the aggregated information is '*inaccurately perceived as the source*' (Flanagin and Metzger, 2008, p.140).
- people who do not follow quality standards, websites that cannot guarantee or filter the information provided and lack of mechanisms to monitor such a sheer volume of information (Flanagin and Metzger, 2008).

In order to understand credibility, Flanagin and Metzger (2008) distinguished two types; the credibility-as-accuracy, which is suitable for evaluating scientific knowledge production, and credibility-as-perception, which refers to the trust someone gains although not being expert. For example, data of OS (GB's NMA) belongs to the first category because of the authoritative way they are produced and of the quality standards they imply, whereas data from Google Earth belongs to the second category because of the majority of people using it without complaining, trusting it and passing the trust to other users. Although Flanagin and Metzger (2008) provide some directions to generally challenge credibility of VGI, their work is basically orientated to the second type of credibility (e.g. user/client rating systems).

Goodchild (2008b, p.242) also mentions the first type of credibility, referring to the way users accept data from an NMA because of the expertise of the producers. Linking it to VGI, he points out that expertise is not always necessary for VGI, since cartographic decisions about map symbols and representation are usually taken by the project developers (who usually design with the help of a cartographer). He further adds that '*Mapping expertise may no longer be one basis on which to judge credibility, or to distinguish the expert from the non-expert*'. An example of this 'locking' of cartographic decisions in VGI is OSM main page (Goodchild, 2008a).

Another aspect of VGI that affects credibility is the fact that in many VGI projects, as well as in other areas of user-created content (like Wikipedia), there is participation inequality; there are very few who contribute the greatest part of information, while the other part is created piece by piece by thousands of other volunteers. Research in the broader area of Web 2.0 by Nielsen (2006) shows that participation on the web can vary from the 'usual' 90:9:1 (meaning that only 1% creates data, 9% edits it – by correcting or enhancing it – and the majority of 90%, called as 'lurkers', simply views or uses it without contributing at all), to worse levels of 94.9:5:0.1 (for weblogs) or 99.8:0.2:0.003 (for Wikipedia). Using a VGI example, in Flickr (2010b) the corresponding figures are Nielsen's 'usual'

90:9:1, while the 1% of active users has contributed the 80% of Flickr's data. Haklay (2007b) argues on the claim of a 'democritised' Web 2.0, since the above numbers could be interpreted as the opposite. As a result, data quality of big chunks of data relies on the credibility of one or a few users.

The fact that in some areas there are many edits can lead to the conclusion that data are of higher credibility and quality (Haklay *et al.*, 2010). Other researchers (O'Reilly, 2005; Bruns, 2008; Budhathoki *et al.*, 2009; Ather, 2009; Basiouka, 2009; Auer and Zipf, 2009) also call upon Linus' Law 'given enough eyeballs, all bugs are shallow', as originally expressed by Raymond (1999, p.29) for open-source projects. When applied to VGI, it means that the bigger the number of contributors for a specific area, the more credible and accurate the contributed data may be.

However, the correlation between contribution and credibility does not apply in cases of geography with ambiguous meaning; when users iteratively change the same features because they cannot agree on the true location or naming, the amount of edits gives the wrong impression of an area that has been cross-checked and reviewed many times. The 'democratisation' of GIS (as used by Kraak and Brown (2001, p.11) who attribute the term to Morisson) can inevitably lead to credibility issues when applying to data creation. When editing someone's contribution is feasible with little or no control over it despite its unknown intentions, it is difficult for someone to trust the spatial data provided (Coote and Rackham, 2008).

Although quality standards and authority barely exist in VGI compared to other spatial data coming from a commercial provider or an NMA, credibility of some projects is relatively high, in some cases similar to or even higher than the information provided from other official sources (Goodchild, 2007a). This refers to the above mentioned credibility-as-perception (Flanagin and Metzger, 2008); people trust VGI projects because they use the provided information for a long time without having been let down or they follow the trust that others show on these projects (Coote and Rackham, 2008; Flanagin & Metzger, 2008).

Devillers *et al.* (2010, p. 396) sum up all the above mentioned VGI characteristics into questions and opportunities stemming from the democratisation of spatial data, generally following a spatial quality direction. For a broader view of VGI, however, other characteristics that are not directly related to data quality are briefly mentioned in Appendix B. These include motivation, ethics and values, the digital divide, copyright issues and sustainability.

2.5. Quality issues of VGI raw data

VGI projects provide various types of data as output. 'Raw data' are considered the input data, collected or uploaded by the contributors, which may require further processing or may directly be presented as the output, depending on the project. Each data input has accuracy limitations, depending on data capture. Additionally, other factors can influence data quality, either on their own, or as a combination with accuracy limitations, leading to heterogeneous data quality. The most important factors are stated next.

2.5.1. GPS technology limitations

In cases were GPS devices are used for data collection, accuracy of GPS devices varies. Usually, the GPS devices used for VGI are single-frequency (L1) handheld devices of relatively low cost. According to their technical specifications, their accuracy now starts from 2-3 meters at best (ESA, 2010). This concerns modern GPS devices with WAAS/EGNOS enabled. However, only three EGNOS satellites cover Europe, and when their signal is blocked by obstacles such as buildings or trees, positional accuracy is reduced to 20 m. The same phenomenon applies in North America, where GPS corrections are provided by WAAS, reaching 3 meters of positional accuracy at best (Garmin, 2010). In case of older or cheaper devices, the accuracy for a stand-alone GPS device starts from 20 m and may decline if the signal reception is not good.

Interference, multipath error and lack of good satellite geometry due to high buildings can reduce the accuracy of the GPS device. Groves (2009, p.44) provides estimations of these errors, for example multipath error could reach 7.4 meters. Geometry of satellites along a thin sky line above our head may increase the error or not provide a GPS fix at all. GPS satellites are constantly moving during the day, so satellite geometry always changes for a specific area, meaning that if mapped later in the same day or next day by the same user and device, the results will be slightly different. As a result, the positioning quality may vary unpredictably.

The way the GPS device is used may lead to less accurate results despite the quality or cost of the device. The density of tracking points will be different if someone is walking, bicycling or driving on a highway, which results in the accuracy of the linear feature represented. This can be evaluated by using tracklogs of different users, but only in cases where there are more than one mapping the same area. Additionally, it is not expected from the amateur users to know how GPS works, e.g. when the GPS signal is strong (thus leading to good results) and when is not, or what may obstruct its vision. Perkins and Dodge (2008, p.27) studied the Manchester Mapping team and realised that

some of the volunteers were walking with the device inside their bag or pocket, which lead to far worse results in positioning.

2.5.2. Attribute incompleteness

Attribute completeness is heterogeneous. There is a significant number of features with no attributes, such as road names, direction, etc. The study cases of this thesis provide such examples. A satellite image does not provide information such as street names, so, unless a user either visits the place or is acquainted with the area, the data will not be complete. Additionally, the degree of mappers' commitment varies. While the street network can be acquired relatively easy by volunteers just by moving around with their GPS turned on, it is more difficult to add information (tags) while walking, cycling or driving. However, enthusiasm or willingness to contribute is not always the case, as there are a lot of GPS devices that do not offer the ability to fill in the feature attributes while moving. Perkins and Dodge (2008) admit that tagging is problematic, time-consuming and less exciting.

2.5.3. Satellite imagery accuracy

Accuracy of the provided satellite images' georeference may also be a matter of consideration. A slightly wrong placement, scale or rotation of the satellite image produces errors that will become part of the vector dataset after the digitisation. Accuracy indications are not given and in any case they obviously vary from image to image. Ortho-rectification errors may mislocate objects by several hundred meters (Ramm *et al.*, 2011, p.49 & p.136). Although in OpenStreetMap (2010o) it is suggested that *'it's fine to assume that the Yahoo imagery is placed accurately'*, it is also mentioned that there are limitations to the accuracy of the satellite images. This can only be checked by '*using multiple GPS readings*' of the same area and realising that they all diverge from the provided satellite image in the same direction and distance, which, however, implies the existence of multiple volunteers with GPS units. In rural areas where such a satellite image problem is likely to exist due to a possible reduced imagery accuracy (e.g. Landsat imagery available), the number of users is unlikely to be high. Searching for 'Google Earth mistakes' on the web, one can find not just a description of them, but even videos describing detailed examples. Goodchild (2007a, p.30) also mentions an error found in Google Earth's imagery in the area of Santa Barbara, reaching 40 m of misregistration and a swath of 60 m width missing from the imagery. He also mentions that the use of this image will inherit the error on VGI output, an error that cannot be found unless Google corrects the mistake in the future, causing in turn the VGI output to look out of place and in need of correction. Haklay and

Weber (2008, p.17) also mention mistakes from aerial imagery, which can be corrected only after ground survey has taken place.

2.5.4. Digitisation in VGI

The way a user digitises differs from person to person. In order to increase the quality, one has to zoom in quite enough. In case of using the maximum available zooming level where the satellite image is visible, the user will have to pan frequently, sometimes waiting for the image to refresh. By zooming out the digitisation process becomes faster, but the accuracy will not be the same. When lines have to be drawn adjacent to other lines or polygon areas, if not zooming enough to see every node of the linear feature, the result will be the creation of 'sliver polygons', unless the user knows how to define some threshold snapping options to avoid such topological errors. There is a trade-off between quality of digitised features and the area covered within a certain time. The user may want to be precise when mapping one's own place, or less accurate when striving to provide faster a larger volume of data so as to reach a higher place in the contribution 'hall of fame'.

2.5.5. Combination of error sources

The combination of all the above mentioned errors leads to the unknown data quality of VGI. It is difficult to estimate where and when each source affects data or where and when an error source enhances, reduces or neutralises the error caused by another source and how the combined error is spatially distributed in VGI.

A way to quantitatively assess data quality of VGI is to ignore the error caused by each source separately and to consider the combined error as one non-systematic and spatially-variable error. By comparing VGI with data of known quality, the combined error can be estimated, regardless of what caused it. After all, in a VGI project there is no real need to know the source of an error, since no user can be held responsible; the actual need is to be able to estimate the error in order to decide if VGI suits for specific purposes.

2.6. Summary

The VGI phenomenon was described, starting from its emergence and its various definitions. Some examples of VGI projects were briefly presented in order to give a better picture of the spatial data provided by volunteers. Specifically OSM was described in more details, as it is used in all case studies. Successively, VGI characteristics were analysed, focusing on those affecting data quality.

While these new spatial sources seem to be promising in terms of their low or no cost, high frequency of updates and increasing coverage, there are other issues that make it difficult to decide their fitness for a specific purpose. Among these are the different numbers of users per area that leads to heterogeneous data coverage, different methods of data capture and lack of standards and metadata that lead to heterogeneous datasets of unknown quality. The increasing data volume, coverage and number of users demand methods that would be appropriate for assessing spatial quality of VGI, so that it can be decided if such a dataset would cover specific demands. Following these, the spatial quality context and its usage for VGI is examined in the next chapter.

Chapter 3: Literature Review

Spatial Data Quality in Geographic Information Science

3. Spatial Data Quality in Geographic Information Science[2]

3.1. Introduction

This chapter focuses on spatial data quality. After introducing the reader to the term, the elements of spatial data quality are analysed, with the aim of identifying the ones applicable to VGI. Methods to measure the quality elements are briefly presented, again focusing on finding the ones (if any) suitable to deal with the different nature of VGI. Finally, a short analysis of previous studies on VGI quality is provided, focusing on how spatial data quality issues were handled.

Although the earliest Geographic Information System (GIS) dates back to the 1960s, concerns on accuracy and uncertainty appeared a decade later (Goodchild, 2002, p.5). As GIS emerged and people started using this technology during the '70s and '80s, they realised that in many cases they did not know how accurate the input data were, or if the quality was sufficient enough for their needs, and even if they had such information, it would not always be applicable to their output data too; very little is known on the way potential sources of error in different datasets will affect the outcome of a GIS procedure which combines these datasets, and this results in a low level of trust in the outputs (Hunter *et al.*, 1995). As a result, during the following decade of '90s, significant research effort with fruitful results led to the engagement of geostatistics to deal with spatial data quality (Goodchild, 2002, p.2), as well as the emergence of standards and tools to assess spatial data quality. Devillers *et al.* (2010) provide a discussion on the achievements, failures and opportunities after thirty years of research on spatial data quality, and link the past with the future by addressing VGI as an opportunity and a source of new research directions.

First of all, a distinction should be made between the terms of uncertainty and data quality. According to Longley *et al.* (2001, p.124-139), uncertainty of spatial data is the inevitable result of our inability to represent the real-world phenomena without a certain minimalism in a database. For example, depending on the level of detail, a building will be represented as a square or rectangle, even if its shape is slightly different or more complicated; as a second example, an arc will often be represented by successive straight lines. If more detail is desired, the database grows bigger, data

[2] Section 3.3.1 has been partially adapted from:

Koukoletsos, T., Haklay, M. and Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS,* [in press - DOI: 10.1111/j.1467-9671.2012.01304.x]

get more difficult to handle and the updating process is more costly and time consuming, so there is always a trade-off between the real world and its representation. This uncertainty affects the three stages of representation, namely the conception of geographic phenomena, their measurement and representation and, finally, their further analysis. For each stage, Longley *et al.* (2001) describe the corresponding sources of uncertainty, including spatial uncertainty, vagueness, ambiguity, scale (for the conception stage), accuracy and error, measurement error, data integration and shared lineage, ambiguity revisited (for the measurement stage), spatial analysis, aggregation and uncertainty (for the final stage of analysis). Data quality is the expression of uncertainty; They consider ambiguity and vagueness as major contributors to data quality, along with error and inaccuracy.

Spatial data quality consists of three parts; the definition of elements of spatial quality, the establishment of metrics to measure these elements and, finally, communication of data quality (Servigne *et al.*, 2006).

3.2. Elements of Spatial Data Quality

The definition of data quality elements has been one of the earliest efforts of researchers. However, no specific list of elements with a corresponding definition is yet agreed. Researchers may use different names for the same elements. Van Oort (2006) provides an examination of data quality elements from five different sources, namely 'Aronoff', 'USA-SDTS', 'ICA', 'CEN TC287' and 'ISO/TC 211'. His work shows the varying definitions, however he manages to assemble them in groups and define ten elements; lineage, positional accuracy, attribute (or semantic) accuracy, logical consistency, completeness, temporal quality (or accuracy), usage-purpose constraints, variation in quality (or data quality scope), meta-quality and resolution. The Federal Geographic Data Committee Standards (FGDC, 1998a, p.1-3) adopts the five elements specified by Spatial Data Transfer Standard (SDTS), which are the first five of the above mentioned.

The most recent attempt of standardising data quality elements is ISO 19113 in 2002 (ISO/TC 211, 2010), which examines spatial quality of a product in comparison with its specification by using the following five elements (p.50):

- **completeness:** presence and absence of features, their attributes and relationships;
- **logical consistency:** degree of adherence to logical rules of data structure, attribution and relationships;
- **positional accuracy:** accuracy of the position of features;
- **temporal accuracy:** accuracy of the temporal attributes and temporal relationships of features;

- **thematic accuracy:** accuracy of quantitative attributes and the correctness of non-quantitative attributes, of the classifications of features and of their relationships.

Additionally, the following three elements are used (when applicable) as indicators of the non-quantitative quality (p.51):

- **purpose** of data
- **usage:** ways to use the data
- **lineage:** description of the source of data, as well as how it was collected and what transformations took place.

However, it is mentioned (ISO/TC 211, 2010, p.50) that in cases that are not addressed in the ISO/TC 211, additional data quality elements can be created.

Advances in technology have made some uncertainty sources and quality elements obsolete, enhanced their importance or changed their meaning. For example, the uncertainty source 'scale' mainly applies to paper maps, while for contemporary digital data 'resolution' is more appropriate (Goodchild, 1993); temporal quality and metadata quality in contemporary and dynamic databases (in terms of updates) gain importance. As a result, since 2002 researchers seek to adapt to the new environment.

Chrisman (2006) discusses the evolution of data quality and how from exclusively dealing with positional accuracy in the beginning, the data quality horizon had to broaden so as to deal with the full information content. He defines the quality aspects as they rose from specific needs in chronological order, namely positional accuracy, attribute accuracy, topology and logical consistency, fitness for use, until he reaches the eras of SDTS and ISO/TC 211 standards, realising that '*the ISO metadata standard and OGC standards rearrange the items in some ways, but retain the same conceptual structure*' as before (p.26), omitting at the same time the 'fitness for use' factor as non-suitable in the producer's perspective approach of ISO Standards. Devillers and Jeansoulin (2006) classify data quality as 'internal' (level of similarity between the actual product and what was expected) and 'external' (level of satisfying the user's needs), providing a very comprehensive example to demonstrate their relationship in a production process. For internal quality they propose the quality elements of ISO/TC 211. External quality is the Chrisman's (2006) omitted fitness-for-use factor, for which they present the following elements (p.40), originally proposed by Bédard and Vallière:

- **Definition:** evaluation of 'what' corresponds to the user's needs
- **Coverage:** evaluation of 'where' and 'when' meets the user's demands
- **Lineage:** evaluation of the source; 'how' and 'why' data are collected and if it complies with the user's needs
- **Precision:** evaluating whether semantic, temporal and spatial accuracy are acceptable for the user
- **Legitimacy:** evaluation of the level of authority, legal and official recognition of data compared to the user's expectations
- **Accessibility:** evaluation of ease of access in terms of cost, time, copyrights, etc

Jakobsson (2002) focuses on how data quality is handled by European NMAs and shows that the five elements of ISO/TC 211 are not used during quality evaluation as much as they should (with usage ranging from 32% to 68%), on the contrary some evaluate different elements; as an example, in Finland's National Land Survey, the quality factors are lineage, completeness, accuracy of geographical location, thematic accuracy and 'currency' (which means how well the dataset meets the required up-to-dateness). Veregin (2005, p.178-184) refers to the five elements of FGDC Standards, but also defines four major categories; accuracy (including spatial, temporal and thematic accuracy), precision or resolution (including spatial, temporal and thematic resolution), consistency and completeness.

Fisher *et al.* (2006) mention six pre-suggested data quality elements, namely lineage, accuracy (positional and attribute), completeness, logical consistency, semantic accuracy and currency. However, they mention that in most cases these terms are not related to research in uncertainty (p.55); spatial auto-correlation of error, vagueness of data or a discordant classification scheme are not addressed by data quality elements. They also add 'precision' as an alternative term to describe data quality. Vauglin (2002) comments further on precision by defining 'geographical precision' and 'geographical resolution', which he distinguishes from digital precision and digital resolution correspondingly by relating them to the geographical meaning of the information represented in the database, instead of a number of digits (precision) or a numeric distance (resolution).

Servigne *et al.* (2006) define and discuss in details eight data quality elements (p.182-185); lineage, geometric (or positional, or spatial) accuracy, semantic (or attribute) accuracy, completeness, logical consistency, temporal accuracy, semantic consistency and 'specific quality' (introduced by France's NMA 'IGN'; it expresses quality-related information not foreseen by the previous criteria, such as

timeliness). Moreover, they add 'precision and accuracy' (aptly describing them with the use of Figure 3.1), 'appraisal and use of quality' (in other words the 'fitness-for-use') and 'meta-quality' (information on the quality, in other words 'quality of quality'). However, they add that quality elements partially overlap each other, sometimes making classification difficult.

Figure 3.1: *Comparison of accuracy and precision (from Servigne et al., 2006, p.184)*

As a final example from a data provider's perspective and regarding vector data, Harding (2006) recognises the quality elements of lineage, currency, positional accuracy, attribute accuracy, logical consistency and completeness. Attention is also paid to understanding the end-user's needs, who relates data quality to fitness-for-purpose.

The above analysis of researchers' description of data quality elements, although far from complete, shows that there is no consensus on a single definition of quality parameters. However, the disagreement can be considered trivial since it reflects differences mainly in classification or significance of the elements, not on their meaning. Further analysis is beyond the scope of this thesis, however there is a need to define which elements are important for VGI.

Since VGI is a new trend in GI Science, there is little that can be found in the literature review regarding how and which spatial data quality elements could also apply to VGI. Coote and Rackham (2008) provide some examples of the five standardised quality elements from the perspective of the VGI user, concluding that VGI quality should mainly focus on Completeness, Consistency, Quality Control (meaning assessment of 'the wisdom of the crowd' and digital 'vandalism', p.11) and Quality Assurance (meaning the statements related to quality and the used methods for assessment). Against their consideration of positional accuracy as of less important, their research finds its measurement significant for the fit-for-purpose assessment as well as for dealing with the lack of

Quality Assurance in VGI projects. Temporal accuracy is explained as attribute consistency for timestamps of creation or deletion of objects, usually hard to be evaluated in VGI due to lack of such temporal data. Logical consistency, finally, is also difficult to be evaluated in most cases due to the lack of any standards or quality assurance procedures.

The nature of VGI (as Chapter 2 described) indicates that some of the ISO/TC 211 data quality elements may require to be slightly customised, so that they could answer the fitness-for-purpose question of generally any VGI linear data source. Specifically:

- **completeness:** presence and absence of features as geometric objects. This would cover cases where there are no attributes or relationships (partially or as a whole) in VGI. (Attribute completeness, however, is separately examined). Data completeness is further divided into omission and commission. Omission refers to missing data that should exist in hypothetically perfect dataset 'A' against which dataset 'B' is compared, while commission refers to excess data, such as recent updates, present in dataset 'A' but not found in the other.
- **logical consistency:** further to the above mentioned ISO/TC 211 definition, logical consistency is a broader area that covers on one hand topological aspects and on the other the validity ranges of values that occur in the data set in spatial, thematic, and temporal parameters (Caprioli *et al.*, 2003). As a result, it consists of topological consistency, temporal consistency and thematic consistency. However, it is quite difficult to find or/and evaluate relevant information in VGI datasets. Coote and Ruckhham (2008) note that there is '*a very marked difference in conceptual approach*' regarding the adherence to data specification, as there are usually no or loose standards and no quality assurance. This renders the measurement of this quality element problematic for VGI.

  Topology, however, needs to be further discussed. It consists of rules and refers to the spatial relationship of objects in a dataset. Topology manages shared geometry, defines and enforces data integrity rules, supports spatial analysis (spatial queries such as the shortest path on a road network, adjacent land properties on a road), supports sophisticated editing (e.g. when moving a common vertice, all objects will be modified) and constructs features from unstructured geometry (e.g. 'spaghetti': constructing polygons from lines) (ESRI, 2005). What is mentioned for logical consistency (regarding the difficulty in having rules, let alone to comply with them in VGI) also applies to topological consistency. This is in agreement with Girres and Touya's (2010) study of OSM in France, as will be discussed in section 3.4. Hence, topology and logical consistency in general is considered to be a quality element that can hardly be evaluated in VGI

generally. For the same reason, it seems not appropriate to form topological constraints that a data matching approach for VGI would effectively rely upon.

- **positional accuracy:** the accuracy of the position of features can be measured by using a reference dataset of higher quality, assuming that the position of the objects in this 'ground truth' dataset is accurately represented. This assumption may not always be true, since all datasets are simplified representations of the real world and they will have some level of uncertainty (Longley *et al.*, 2001, p.124-139), however it is a compromise usually followed in the literature for quality analysis (examples will be mentioned later on).
- **temporal accuracy:** Coote and Ruckhham (2008) present some examples to show that information about the temporal attributes and temporal relationships of features is hard to find or quite fuzzy in VGI. Additionally, it will not be generally applicable in any VGI case. This thesis will not evaluate this quality element, however it can be indirectly negated when comparing datasets acquired at the same time. Thus, the chances that temporal accuracy will be generally similar between the datasets are higher.
- **thematic accuracy:** the above mentioned ISO/TC 211 definition needs some modification to cover generally all different VGI sources. To cover general cases of VGI, this should be limited to the attribute accuracy evaluation of appropriately selected data fields. Attribute completeness needs also to be examined, along with attribute accuracy rather than during data completeness evaluation. This approach seems more appropriate for VGI, because some features may exist as objects but may not have attributes.

The next section moves deeper into spatial data quality by discussing methods of data matching. This is essential before any further quality analysis that relies on datasets' comparison, so as to ensure that corresponding objects are examined. Successively, methods to measure data completeness, attribute accuracy and positional accuracy are discussed.

3.3. Establishment of metrics to measure data quality elements

Data quality measurement is addressed in a number of ways, depending on the element measured. Some researchers evaluate data quality in terms of risk (e.g. Agumya and Hunter, 2002), others implement stochastic models (e.g. de Bruin, 2008), and others use more geometrical-based methods, which address a specific type of entity (e.g. point, line or polygon) and by expressing their result in measurable map units they assess its positional accuracy (e.g. Goodchild and Hunter, 1997).

For the measurement of the selected quality elements in the context of this research, the quality of a VGI dataset can be assessed by comparing it with another dataset, which is assumed of higher quality. Section 1.4 mentioned the importance of linear data type in spatial datasets, usually used to represent networks, which will be the data type concerning this thesis. In the next section, methodologies that can be applied to measure quality of linear datasets will be examined, among which the most appropriate one (or combination of methods) will be further explored.

Before continuing, there is a need to explain the terms that will be used to describe the linear data. A linear dataset consists of simple or complex lines, called polylines. Every such object is one record in the dataset with a unique id number, called a 'feature'. A feature could be a simple line defined by two points or a complex polyline with many vertices. Features start and end at points that could be junctions (a meeting point for many features – road intersection for road networks) or end-points (dead-ends). A feature can be further divided into 'segments', which are straight lines that begin and end at successive vertices of the polyline that forms the feature. Segmentation may need to be performed for reasons of data handling, however the basic object unit of any dataset remains the feature, not the segment. A 'road' consists of one or more successive features with the same name attribute (if available). A feature can be further split into one or more successive segments. Each feature has compulsory geometric attributes and a unique identifier (id). Additionally it may have thematic attributes such as name, road type, alternative name, maximum speed, number of lanes, etc. Features that form a road will normally have some common thematic attributes (such as the road name and road type). Segments that belong to a feature inherit the feature's thematic attributes and id.

3.3.1. The necessity for data matching

Data between two different datasets, yet for the same area can be different due to:

- Different data collected, usually based on the provider's objective. Data of low commercial value and high collection cost may not be present in official or commercial datasets. As an example, cycleways, bridleways, steps or footpaths are collected by OSM users but not by OS, GB's NMA, for their most detailed MasterMap dataset. Different density of data also leads to inconsistencies between datasets.
- Different methods of collection, different data structures. A feature in a dataset aimed for routing purposes will usually start and end at a junction, while on another dataset the same feature may be represented by more than one features, or it may be a small part of a bigger

one that corresponds to many features in the first dataset. Mustière and Devogele (2008) provide such examples.

- Different sources of raw data. As an example, when using a generalised map as a source to produce vector data, detailed information not included in the source will also be absent from the final product. Scale also leads to different representations; a round-about for example can be a polygon, a circular (linear) object or a junction (point), depending on the scale.
- Different timing of updates, which may lead to the presence of relatively new data only in one of the two datasets.

Data matching is usually the first step in a data quality analysis and it is often considered as part of the data preparation stage. There are basically two options of performing data matching. A manual approach relies on an operator's competence and is usually very efficient for small areas. In bigger areas or denser data, however, the time and effort increases dramatically, human errors or negligence are more likely to occur and any repetition of the data matching process seems extremely tedious. Automatic data matching, on the other hand, can be faster for larger areas, however its efficiency depends on the nature of the data and the way the designed rules apply to the data involved. However, hybrid approaches also exist (semi-automatic data matching).

Data matching is traditionally associated with 'conflation' or 'data fusion', the process of combining different spatial datasets to enhance one of them or to create a new integrated dataset. Conflation research in GIS dates back to the 1980s and includes two general stages. The first is finding corresponding objects (data matching) and is the one related to this study, while the second deals with transformation and integration of the datasets, not addressed here.

There is no largely accepted and applied matching method for linear datasets. According to the nature and scope of the datasets involved, different factors must be taken into consideration. Doytsher *et al.* (2001) provide a description of matching algorithms for linear datasets developed from the middle '80s and forth. They distinguish them into point-based and line-based methods, mentioning the advantages of the latter due to the linear nature of most of the map data. They propose a line-based algorithm, using geometric and topologic constraints. However, using topology when information is missing from one dataset can be problematic (Safra *et al.*, 2006), which is the case of VGI. The latter ones propose a point-based method, however they examine nodes relying on semantics and do not use any other non-spatial attributes of the linear datasets.

According to Walter and Fritsch (2001), the matching approach can be 'feature-based', when geometry and / or attributes are examined, or 'relation-based', when relations between objects are considered. The choice depends on the nature of the data. They present an automated relational matching method for integration of linear spatial data from different sources. They limit their constraints to geometric ones to identify possible matching pairs, which they further evaluate using measures from information theory and statistics. They mention that their approach, although automated, is significantly less successful for cases of completely different data in certain areas (due to temporal difference), which makes their method rather unsuitable for VGI.

Devogele *et al*. (1996) provide a combination of relation-based and feature-based approach, to deal with data matching between different scales. Scale-transition relationships are described between classes and types. An integration technique transcribes the relationships and creates a multi-scale schema. Data matching follows, using semantic, topologic and geometric constraints. There are three stages to examine roads (using semantic constraints), crossroads (using geometric and topological constraints) and sections (using semantic and geometric constraints). Using Hausdorff distance (discussed in section 3.3.4), sections are classified as 'matched', 'unmatched' or 'litigious' (when no homologous sections can be chosen). Their tests provided perfect matching results for roads, good for '1-1' crossroads and average for '1-many' crossroads. They focus, however, on data from different scales and generalised datasets. Additionally, relying so much on data classes may not be applicable for VGI due to the unknown, non-existent or non-standardised classification by VGI contributors, which will render data matching problematic (Girres and Touya, 2010).

Dunkars (2003) distinguishes data matching into matching of database schemas and matching of actual data, however, along with Mantel and Lipeck (2004), they also focus on generalisation and deal with data acquired by the same organisation but represented in different scales.

Mustière and Devogele (2008) propose a method for network data matching between datasets with different levels of detail. They focus mainly on geometry, seeking for node and arc matching based on Hausdorff distance (described in section 3.3.4). As they claim, their approach is not very efficient when data are inconsistent or complex and does not take into consideration non-spatial attributes such as road names.

Safra *et al.* (2010) tackle data matching in heterogeneous sources, examining location-based joins through sequential or holistic approaches. They present a point-based approach using only locations

of objects, however they assume that in both datasets distinct objects represent distinct real-world entities and that accuracy is uniform for each dataset, which does not always happen in VGI.

Gabay and Doytsher (2000) present a linear matching process that relies on directional and positional proximity and targets large scale engineering maps. They first examine points and segments, using buffers for proximity and an angular tolerance for direction, and then they match features based on sequentially matched segments. However, although it applies to data derived from large scale maps with topological differences, it is assumed that they are topologically organised and with homogeneous accuracy, which makes it inappropriate for VGI, while they do not include thematic attributes in their analysis.

So far, all the mentioned studies dealt with homogenous and standardised data. On the other side, previous studies on VGI quality (Cipeluch *et al.*, 2010; Girres and Touya, 2010; Haklay, 2010c) evaluate VGI by comparing it with official or proprietary datasets of known quality. However, their approaches are difficult to replicate and scale due to the need of manually matching features between the two datasets. An automated matching procedure is offered by Ludwig *et al.* (2010) for OSM in Germany, however their method is specifically designed for geomarketing purposes and for the datasets involved, so they exclude roads of no business interest (such as motorways or roads with no name attribute), rendering their method unsuitable for the general case of VGI.

Generally, a matching procedure for linear datasets could rely either on geometry or on thematic attributes of datasets involved, or on a combination of both. Among the geometric constraints that could be used for linear datasets, the most obvious is distance; corresponding objects between two datasets are sought within an appropriate distance. This will reduce significantly the number of possible matches for a feature, however it will usually not be enough or correct; close to junctions there can be a lot of possible matches within a specified distance, especially in urban areas. Another geometric factor could be the feature's shape or length. In most cases, however, datasets come from different sources and the same linear object can be represented differently, using lines of different length, polylines with different number and position of vertices or even more than one features. Finally, directional comparison could also be performed. De Smith *et al.* (2009) however find it problematic for three reasons. The first two refer to the way one line is represented, the extent of generalisation and the lack of knowing the true start and end of the line to perform a directional analysis. The third reason is the orientation of the features, usually defined during the

capture procedure according to the succession of points, and the difficulty in finding a mean direction for all the parts that form a polyline.

Moving away from geometry, looking for a match regarding the feature attributes is also possible, but in VGI case this can be far more unreliable due to the sporadic and unpredicted lack of attributes. The most appropriate attribute to compare in a road network would be the road name, taking into consideration misspelling or abbreviations used for the same object.

None of the presented methods can efficiently deal with VGI nature. Data matching should be designed differently for such data, accepting VGI heterogeneity, combining geometry with attribute constraints in a complementary way. Thus, corresponding data with missing thematic attributes could be matched based on their geometry, while corresponding data that fail to comply with the geometric constraints could be found as matching based on the same road name attribute. Such an automated method is proposed in the next chapter.

3.3.2. Assessing data completeness

Data completeness refers to the presence or absence of features in one dataset, along with the presence or absence of their attributes and relationships (ISO/TC 211, 2010). This presence-absence consideration assumes a complete reference dataset, as well as a comparison of similar types of information. Lack of data that should be present in a dataset according to its specifications is referred as omission, while excess data that should not be present (again according to the specifications) is referred as commission (Servigne *et al.*, 2006; Coote and Rackham, 2008). Servigne *et al.* (2006) describe two types of completeness: model completeness, which refers to an evaluation in terms of fitness-for-use, and data completeness, which refers to measurable errors regardless of the application. Data completeness is further divided into formal completeness, which examines the adherence to the data structure and standards, and object completeness, which examines the existence of entities or features. The latter is followed by attribute completeness. Due to the nature of VGI, however, the lack of thematic attributes which may occur in some areas (Maué and Schade, 2008) demand new approaches for assessing VGI completeness. This thesis examines the existence of corresponding objects to assess data completeness, while attribute completeness is examined separately as attribute accuracy.

When comparing datasets of different structure and information, ‘completeness’ is difficult to define. Both datasets may contain information not present in the other dataset due to different

specifications and objectives, which may in turn imply collection of different information. VGI is usually among these cases. A better term in such cases might be 'data agreement' instead of 'data completeness'.

Following the data matching process, data completeness could be easily estimated by calculating the amount of data present in both datasets (matched or common data) compared to the original dataset's size. Since the number of features for the same object representation may differ between two linear datasets, calculating the dataset lengths (matched, non-matched, total) will give the necessary information. A proper application and handling of data matching is proposed in the next chapter, so that the 'data agreement' meaning moves closer to 'data completeness'.

3.3.3. Assessing attribute accuracy

Thematic or attribute accuracy evaluation depends on the data type, which can be quantitative (e.g. precipitation, usually expressed in numeric format) or qualitative data (e.g. road types, road names).

Van Oort (2006) and Servigne *et al.* (2006) mention four measurement scales for attribute completeness: Ratio, Interval, Ordinal and Nominal. The choice depends on the type of data. While qualitative (or numeric) data can easily be compared, quantitative (or text) data are more difficult to handle. The Nominal scale used in such cases (e.g. road names, road types or land cover) is a suitable unordered scale, however, possible errors are traditionally considered as misclassification, meaning that there is a well-defined range of values and that a wrong value is selected. This is assessed by sampling the data to check for the correct classification (Caprioli *et al.*, 2003). Servigne *et al.* (2006) and Devillers and Jeansoulin (2006) also mention classification when it comes to precision of non-spatial attributes. This, however, assumes that all objects have attributes and additionally values of a limited range, which is not always the case for VGI; it can lead to inconsistent classification schemes and the use of standard confusion matrices [3] cannot be applied (Al-Bakri & Fairbairn, 2010). As an example from the first case study of this thesis, there are nine road types used for Greater London by OS, while for the same area the corresponding VGI dataset from OSM includes more than 100, among which there are misspelled, abbreviated or incompatible data entries.

A different approach is necessary for VGI. Misspelling or use of different abbreviations should be accepted as correct when it is obvious that they refer to the same object. Assuming again that the

[3] A confusion matrix (Kohavi and Provost, 1998) is a square matrix, used in classification systems, which contains information about the actual and predicted classifications. In its simplest form it is a 2x2 matrix that informs about true positives, false positives, true negatives, false negatives.

reference dataset is complete and correct, this can be achieved by functions or algorithms that compare the two strings. Figure 3.2 presents the possible outcome of such a comparison.

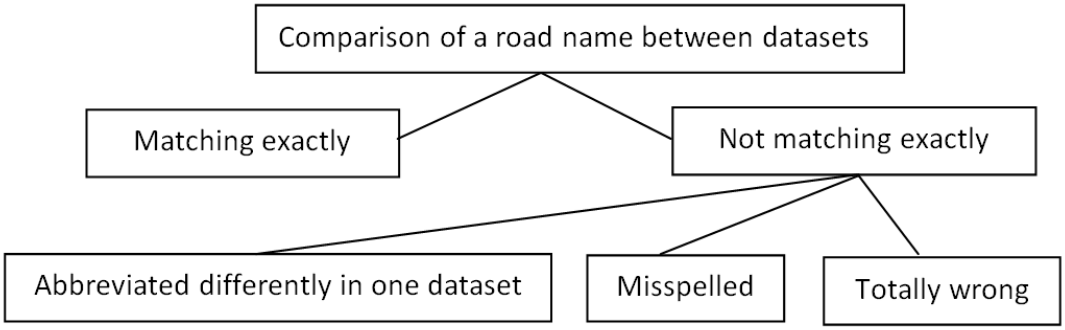

Figure 3.2: *Comparing thematic attributes*

The 'abbreviations' problem can be dealt by manually collecting all possible abbreviations and create the code accordingly to take them into account. However, new abbreviations in a different area, country or in the future would demand reprogramming. The OpenStreetMap (2011) wiki provides extensive lists of abbreviations for a number of languages, which demonstrates the complexity of the use of an abbreviation index, especially in countries where more than one language exist. As for misspelled names, there is not always a clear line between the misspelled and totally wrong version. There are ambiguous road names, widely accepted in more than one version. An example is the use of apostrophe, i.e. is 'Queen's Gate' similar to 'Queens Gate' or 'Queens' Gate'? There is a lot of debate over the apostrophe use (The Turner Ink blog, 2009; The Telegraph, 2009). Other languages can have similar problems as well. On the other hand, when VGI is examined, the evaluator should be more lenient and accept a slightly misspelled road name as an accurate one, because it is either local knowledge that may be more up-to-date, or a result of the user's reduced grammar knowledge. The problem then lies in defining the threshold between what leads to the same meaning as the official road name (e.g. 'Lilly Road' with 'Lily Road') and what not (e.g. 'Lilly Road' with 'Lullaby Road').

Various algorithms exist for text comparison. A brief description (PHP online manual, 2011a; Charras and Lecroq, 1998): 'Soundex' supposes knowledge of the pronunciation but not the spelling, therefore relies on a key created by words pronounced similarly. 'Metaphone' is a similar and more accurate algorithm, which uses the basic rules of English pronunciation. These need to be customised for different languages. The 'Levenshtein' algorithm measures similarity between two strings by calculating the least number of edits that are needed to modify one string to another. Abbreviations, however, are not covered in any of these functions. PHP's 'similar_text' function is a

much simpler and faster one, which returns the number of similar characters between the two strings. Most of the above algorithms are supported by programming languages, such as PHP (PHP online manual, 2011b). The first two can be customised to be applied to languages other than English, while for the last two it is not necessary. Among them, similar_text seems to be the simplest and fastest in terms of processing, and under some conditions it could cover cases of misspelling and abbreviations at the same time. For these reasons, it is considered as most suitable for this thesis.

3.3.4. Assessing positional accuracy of point and linear features

Point positional accuracy can be an easy task, simply by comparing its coordinates with those of the corresponding point of the reference dataset. The result can be a root mean square distance or percentiles of the distance distribution (which is usually considered to be the Gaussian Normal distribution) (FGDC 1998a; FGDC 1998b; Leung and Yan, 1998; Tveite and Langaas, 1999; Veregin, 2000; Zandbergen 2008). However, difficulty lies in defining what to compare, not how to do it; data preparation should ensure that points exist in both datasets and that the same semantic points are selected to be compared. This can be achieved by using a suitable algorithm (Safra *et al.*, 2010).

Unlike point features, for which spatial positional accuracy standards already exist (FGDC 1998b), linear features cannot be easily compared and there is no standardised method, despite the fact that they usually are the majority of objects on a map, as already mentioned. However, the literature provides some suggestions, and different approaches continue to appear. Some of them will be examined here.

One way to compare lines or polygons is through the positions of their boundary points. However, although a line is defined by its start and end point, it consists of infinite points between them, and methods that compare only the start and end point of a line neglect all the other in-between points, paying no attention to the shape and curvature of the line.

Another possible approach, followed by quality standards published by FGDC, is to compare well-defined and adequately distributed points. These points are often called ‘primitive’, and researchers propose various solutions to carry out such a comparison (Kiiveri, 1997; Leung and Yan, 1998; van Niel and McVicar, 2002). However, these approaches simply make the sample of points richer. Furthermore, it is not always possible to identify well-defined points in a road network apart from the road intersections; in rural places intersections are far fewer and in any case, there is still no information for the in-between part of the line. Apart from that, the error distribution between well-

defined points varies due to generalisation or different accuracy when obtained, and may well be different from the distribution error of the linear part that links the points (Goodchild and Hunter, 1997, p.300). As a result, it does not seem to be a good choice for this thesis.

Another approach is the use of the epsilon band. This has been used for quality evaluation (Chrishman, 1989; Shi, 1998), as well as for linear generalisation in cartography (Nakos *et al.*, 2008). According to this method, a fixed buffer is applied around the reference line that will include the whole part of the tested line. The buffer is created by a disk of diameter epsilon rolling along both sides of the line. However, Goodchild and Hunter (1997, p.300) note that this method is not robust enough because it is very sensitive to outliers and thus it depends on the sample. For example, if there is a gross error in a specific area, it will affect the buffer width and the results would be different if this area was not included. Since the need for this research is to implement a method in a relatively large area, such errors cannot be predicted and will probably exist in any VGI dataset. Thus, this method is not suitable either.

Another method is to take sample points along the tested line and to measure the shortest distance to the reference line. This, however, does not take into consideration a possible distortion of individual points, and the calculation of the shortest distance is the most optimistic approach, since the true distortion can be bigger (Goodchild and Hunter, 1997; van Niel and McVicar, 2002). An example can be seen in Figure 3.3, where the shortest distance is the continuous line and the distorted distance is the dashed one.

Figure 3.3: *Case of distortion of individual points (from van Niel and McVicar, 2002, p.459)*

Hausdorff distance is another method that is mostly used for automatic matching of linear datasets (Mustière and Devogele, 2008). It is the maximal distance between two lines L1 and L2, which equals to the maximum of distances from any point of one of the lines to the other. Ariza-López *et al.* (2011) compare it with three other line-based positional accuracy methods and find that it produces non-representative estimations when dealing with datasets with different data density, which

implies the existence of non-corresponding objects. As a result, it does not seem quite efficient for VGI.

Goodchild and Hunter (1997) propose a method that extends the epsilon band method, removing the drawbacks mentioned above. A buffer of increasing width (instead of fixed) is created around the reference line, covering accordingly an increasing percentage of the tested line (Figure 3.4). Starting from a minimum width buffer, the process is iterative and each time the length of the tested line inside the buffer is calculated, until the desired percentage of coverage is reached. For example, if 95% of the tested line needs to be inside the buffer, the process begins with a minimum buffer width and modifies it (by increasing or decreasing it) at every step. When the percentage is reached, it can be concluded that the accuracy of the tested dataset equals to the buffer width for the specific data percentage. The advantage of this method is that it is relatively insensitive to outliers. For example, a gross error in a feature's position will be part of the percentage not covered by the buffer and will not affect the buffer width. Additionally, according to Hunter (1999, p.27), the method is statistically based and does not require matching points between the datasets. This seems to be a good choice as a method for this research. A simplified version of the Increasing Buffer Method (IBM), as it will be referred to from now on, is to use one (or more) pre-defined buffer values on the reference dataset and calculate the overlap percentage accordingly (e.g. overlap percentage for buffer width 8 m), without having an iterative process.

Figure 3.4: *Increasing buffer method (from Goodchild and Hunter,1997, p.301)*

Tveite and Langaas (1999) present a Buffer Overlay Statistics (BOS) method, which seems to be an extension of Goodchild and Hunter's (1997) buffering method. In fact, they originally proposed their method back in 1995 (Tveite and Langaas, 1995), however without implementing it or presenting any examples. They refined it later, taking Goodchild and Hunter's (1997) method into consideration. The buffer analysis is carried out in both the tested and the reference line. This results in four different polygon areas (see Figure 3.5), and conclusions can be made by calculating statistics in these areas (such as total area, number of polygons, total perimeter). Their proposal addresses also completeness and miscoding of linear datasets; the advantage of choosing to buffer also the tested line is that incompleteness of the reference dataset, in comparison to the tested one, is also considered. Such an approach is more suitable when no dataset can be regarded as of higher quality, which is not the case of this thesis. Additionally, data completeness, as addressed in this study, is likely to give more accurate results, since they are calculated after comparing the objects themselves and not based on a buffered area that may also contain other linear objects or fractions of them. Ariza-López *et al.* (2011, p.712) describe this as '*undesirable behaviour of the method and bad distance or inclusion estimation*'.

Figure 3.5: *Polygon areas after applying buffer in both lines (from Tveite and Langaas, 1999, p.33)*

Another extension of Goodchild and Hunter's (1997) buffering method is proposed by Heo *et al.* (2008). Their research takes into consideration that the tested line may have an offset. As they describe, the increasing buffer is applied on the reference line and the overlap percentage of the tested line is calculated. An estimation of the mean and standard deviation follows. A test was carried out using different fixed values of offset of the original line, and the results showed that by using two parameters (mean and standard deviation) instead of one leads to lower values of standard deviation and RMSE (Root Mean Square Error). However, this method cannot be applicable here; an offset in VGI datasets may locally exist, but it will be unpredictable, mostly as a result of GPS

data affected by a combination of factors that influence the quality of positioning, mentioned in section 2.5.1. As a problem, it will only affect a minor part of the studied area. Additionally, to find these erroneous GPS data would require splitting the datasets according to the user and date of data collection, and examine patterns of possible offset for specific features. On the other hand, geo-referencing of the satellite imagery can lead to the same errors in digitisation and apparently to a similar data shift. However, even in this case, it will only affect digitised data, which may lie next to correct data, differently captured. Although it could be a good way to correct VGI in a semi-automatic way, it goes beyond the scope of this thesis.

Ramirez and Ali (2003) examine the Bias factor (which is a comparison between two lines and calculation – analysis of the length falling on each side of the reference line), the Distortion factor (which is a comparison between the standardised parameterization of two lines), the Fuzziness factor (which refers only to the end points of the linear segments to be compared and relates to their definition and identification), and the Generalisation factor (which assumes that one linear dataset is generalised). They conclude that these factors are linearly independent.

Finally, Ariza-López *et al.* (2011) compare four already mentioned line-based positional accuracy methods: Mean Distance Method (based on the epsilon band), Hausdorff distance, Goodchild and Hunter's (1997) IBM, and Tveite and Langaas' (1999) BOS method. Although their basic concern is the sample size, useful information is provided for the above methods. Regarding the two buffer methods, the IBM, tested in its simplified version, shows a better performance regarding the sample size and gives better estimations than the BOS method. This strengthens the justification of using the IBM in this thesis.

3.4. Previous research on VGI quality

One type of research conducted so far on VGI mainly focuses on a quantitative analysis of data quality, which is carried out by comparing VGI with datasets of higher accuracy for selected areas of limited size. This type of evaluation has been carried out in various countries such as UK (Basiouka, 2009; Ather, 2009; Haklay, 2010c; Koukoletsos *et al.*, 2012), Ireland (Cipeluch *et al.*, 2010), Germany (Ludwig *et al.*, 2010; Zielstra and Zipf, 2010), France (Girres and Touya, 2010), Greece (Kounadi, 2009), Switzerland (Ueberschlag, 2010). Since access to VGI vector datasets is essential for such a comparison, most of these studies refer to OSM, which is so far the only VGI project that allows free access and data downloading for the above.

Haklay (2010c) compares VGI (OSM) with five 1:10,000 raster maps from Ordnance Survey, as well as VGI (OSM) with a generalised official dataset (Meridian) for the Motorways of England. He focuses on positional accuracy and completeness by using visual and statistical ways of comparison respectively. In the second non-manual case, data matching is avoided by comparing one specific road type and assuming that road classification in VGI is correct, inevitably ignoring VGI non-tagged or erroneously tagged features. For positional accuracy he follows the IBM method of Goodchild and Hunter (1997) (presented in section 3.3.4) in its simplified version, using a pre-defined buffer value to calculate the corresponding overlap percentage. Some manual pre-processing was performed to ensure similar representation of roads. For data completeness he calculates and compares the total length per km^2, according to an appropriately created national grid, excluding coastal cells. He further studies social justice and equality of the collected data by combining data volume with the number of users and the UK government's Index of Deprivation 2007. His results show that rural areas are not as well covered as urban ones, while there is also evidence of lower quality in socially marginalised areas. Although the reference dataset he used is a generalised one, he concludes that it cannot be replaced by OSM.

Ather (2009) extends Haklay's initial study in London by applying it on A and B Roads apart from Motorways and using the most accurate official dataset available for selected areas of limited size (25 km^2). Data matching between datasets is performed manually. Positional accuracy is again based on the simplified version of IBM, applying different buffer value to each of the examined road types, based on road width assumptions. Since the simplified version calculates the overlap percentage (VGI percentage inside the buffered reference dataset) instead of the buffer width, it is mentioned that it is not suitable to provide an exact measure of accuracy. Road name attribute had to be edited manually for corresponding objects before beginning the buffering process. Miss-classification in VGI road types is also encountered, though it is corrected manually. Attribute completeness is assessed by calculating the total length of data with road name attribute. Assuming that lack of name attribute means data captured from satellite imagery, he further expected lower positional accuracy for this data, which is proven true for the areas tested. Additionally, user analysis is carried out to test Linus' law 'given enough eyeballs, all bugs are shallow', proving that in most of the tested areas an increased number of users lead to data of better attribute completeness, but not necessarily of better positional accuracy. Basiouka (2009) follows the same methodology as Ather (2009).

Kounadi (2009) evaluates VGI in an area of 25 km^2 in Athens, Greece. She performs data matching manually. Datasets are analysed tile by tile using a grid of 1 km^2. Length is used to assess data

completeness for the whole dataset, while the evaluation of other quality elements is applied on three selected road types. Miss-classification of road types in VGI is handled manually. For attribute completeness, the number of distinct road names is used, however, due to the different data structure, additional editing was necessary (e.g. grouping features by name). For attribute accuracy the road length is calculated manually. Topology problems are also solved manually. For positional accuracy, the simplified version of IBM (mentioned before) is used, applying specific buffers to each feature and calculating the average overlap per tile. Results show that data completeness is relatively high in the tested area, attribute completeness is quite low, attribute accuracy is high and positional accuracy good, while she also mentions the difficulties in finding corresponding road types due to the different classification between the datasets.

Cipeluch *et al.* (2010) compares VGI with proprietary datasets in Ireland (specifically OSM with Google Maps and Bing Maps), to study completeness, currency and accuracy of the data. Their method is manual and relies on visual comparison. They also assume that VGI road type classification is correct and they apply their analysis to selected OSM road types.

Zielstra and Zipf (2010) evaluate data completeness of OSM in Germany by comparing it with a proprietary dataset (TeleAtlas). They also deal with heterogeneity by splitting data into tiles of 1 km^2. However, their results are based on calculating the total length of the datasets for each tile without previously matching the data, so a similar length for a tile may not necessarily mean that VGI is complete, as it may include additional information not present in the reference dataset while missing other data of similar length at the same time.

Ueberschlag (2010) studies VGI quality in Switzerland by comparing OSM with official and commercial datasets for the Geneva canton. Her approach is not fully automated, however she tackles many quality aspects. For positional accuracy she follows the simplified IBM version, calculating the percentage of VGI dataset within a specific distance of the reference dataset, however she applies it only on sample roads. She also deals with attribute completeness by counting the number of named segments, and with attribute accuracy by comparing the number of road names between the datasets. However, as her results prove, the number of named segments is not a useful indicator, as it relies on data capture, and a low accuracy of OSM was partially the reason of looking only for exact road name matching between the datasets. Additionally, this cannot cover cases where attribute is missing from one or more segments that comprise a road. Before the quality analysis, no data matching is performed; instead, selected OSM road types are rejected, assuming a

correct classification in the VGI dataset. She further moves on evaluating user equality by counting the number of users per area and comparing it to the population density. Finally, she applies visual comparison using orthophotos as an alternative way to assess data completeness and accuracy.

Girres and Touya (2010) study VGI in France by comparing OSM with datasets from 'Institut Géographique National' (IGN), the NMA of France. They target point, linear and polygon objects, however their method is applied on sample areas and data only. They deal with data matching by manually selecting 'homologous' objects. For the positional accuracy of linear objects they use the Hausdorff and the average distance. For attribute accuracy they compare strings using the Levenshtein distance, which also covers misspelling in VGI. Their findings also show that quantitatively attribute quality gets better where the number of contributors is increased. For semantic accuracy they examine the road type correspondence between the datasets and they find it problematic because of the different classes used. For data completeness they compare the number of objects and the total length between the datasets for the whole country, however they mention that their method is not as systematic as Haklay's (2010c). For logical consistency they examine intra-theme and inter-theme consistency, concluding that lack of standards and integrity constraints results in difficulties in measuring VGI logical consistency because of topology errors. For temporal accuracy they compare the mean capture and version date with the number of contributors. Based on their results, they also examine usage, further discussing the suitability of OSM for navigation, automatic generalisation or urban planning in France. They find it unsuitable due to heterogeneity, problematic logical consistency, incompleteness and low topological consistency.

Ludwig *et al.* (2010) study VGI in Germany by comparing OSM with proprietary datasets from NAVTEQ. This is the only study so far to offer an automated method of evaluation, which allows repetition in different areas or when data are updated. They assume that road type classification and name attribute is correct in both datasets. They further assume that a road object in OSM is at least of the same length or longer than its corresponding one in NAVTEQ dataset. They are interested in populated roads, so they ignore all roads with no name attributes in the reference dataset, along with Motorways and their distributor roads. Data matching combines geometric and thematic constraints and is performed using three different buffers (5,10,30 m) around reference objects. It is not very efficient for VGI objects with a distance bigger than 5m from the corresponding reference object (as a result of not mapping the axis of the road) or for those with no name attribute. For data completeness they count the number of unmatched objects, yet they cannot

determine VGI over-completeness because of the previously rejected information. For attribute accuracy they examine four attributes, while they also compare primary and secondary road names during the matching process using the Levenshtein distance. For positional accuracy they use the same buffers from the data matching stage, which gives results similarly to the simplified version of the IBM.

Al-Bakri and Fairbairn (2010) test VGI and official datasets against field survey data, specifically OSM and OS's MasterMap in small areas of Northumberland, in order to assess positional and shape quality. They use three methods; the point sampling method, the simplified IBM version for some of the linear data only, and the area shape measure. They intend to assess possible data integration, however their results cannot be generalised because of the heterogeneity of VGI and the need for field work.

Haklay *et al.* (2010) test VGI credibility against the number of users. Their results prove that having more users contributing in an area generally leads to better positional accuracy, which is in agreement with Linus' Law. Their tests show that positional accuracy significantly increases when having from 5 to 13 users, above which it remains level and below 6 m. Based on their results, they suggest that the number of users could indicate positional accuracy without the need of a reference dataset. Although they also check the correlation between data completeness and socioeconomic factors, they conclude suggesting that the number of users has to be examined against data completeness more systematically, as well as against attribute accuracy.

Other types of research on VGI quality do not require comparison with datasets of known quality. For example, research on VGI metadata has been conducted by Antoniou *et al.* (2010), who addressed data quality issues from the perspective of wiki-behaviour of the OSM project; after realising that democracy results in too many tags for the same category and a reduction of quality, they propose an XML schema to model tagging according to OSM's proposed feature list, which could later be used to track violations of the proposed rules.

Van Exel *et al.* (2010) mention that the methods for traditional quality elements assume homogeneous and consistent datasets, which make them unsuitable for VGI. Instead, they introduce the concept of 'Crowd Quality' to describe and quantify VGI quality. Their theoretical approach has two dimensions, the 'User' quality, determined by local knowledge, experience and recognition, and the 'feature' quality, which is expressed and assessed by the traditional quality elements.

Mooney *et al.* (2010) address VGI tagging from the ontology side. They count the number of objects that have been annotated by the users in respect to special and verifiable tags such as the source, description, attribution and source-url. However, they simply compare usage of these metadata between different countries, so they do not actually measure attribute accuracy.

Research has also been conducted on VGI on other areas with objectives irrelevant to data quality. Mummidi and Krumm (2008) present a technique for data mining of POIs. O'Brien (2009) proposes a technique to automatically produce maps from a VGI database for the use of the navigational sport called 'street orienteering'. Schmitz *et al.* (2008) present an example of a more sophisticated use of OSM from OpenRouteService. Auer and Zipf (2009) argue on how VGI geodata and Open Standards could fit together.

These were just some examples of research on VGI. Concluding, each VGI research dealt with different areas, for different purposes and using different methods (although some may have something in common). Section 4.2 will refer to their limitations that leave gaps in the literature, which this thesis aims to cover by providing an appropriate quality analysis framework.

3.5. Summary

Spatial data quality in GIS was examined, analysing the spatial quality elements that need to be measured in order to describe it. Slightly different definitions and approaches are given in the literature and, although standardised, the nature of data under examination may differentiate the importance and scope of these elements. For the VGI case, data completeness, attribute and positional accuracy are considered as the most representative quality elements in this context. For each of them, various measurement techniques were briefly described, in order to select the most suitable ones for this study. Previous research on VGI quality used different ways to measure it, however there is plenty of space for improvement.

For data matching, a procedure that is necessary to precede the quality analysis so as to ensure that corresponding objects between the datasets are compared, the provided automated methods do not seem suitable for linear VGI. Some of them rely solely on the geometry of features (Gabay and Doytsher, 2000; Doytsher *et al.*, 2001; Walter and Fritsch, 2001), while by combining geometry and appropriate thematic attributes it probably leads to a more accurate matching, especially in cases for a dense network where geometric constraints may not be enough. Other methods assume a uniform positional accuracy (Gabay and Doytsher, 2000; Safra *et al.*, 2010), which is not the case of VGI. Some of them prove to be inefficient when data are inconsistent or complex (Mustière and

Devogele, 2008), not updated in one dataset or different (Walter and Fritsch, 2001). The use of topology (Doytsher *et al.*, 2001) can also be problematic due to the lack of such information in the VGI source (Safra *et al.*, 2006; Girres and Touya, 2010). As a result, a new method needs to be developed for VGI, which will be able to monitor its heterogeneity, random lack of thematic attributes and topology, variable density and positional accuracy. Additionally, it should be developed in a way that will aid the measurement of quality elements afterwards.

An efficient data matching method leads to a more accurate measurement of data completeness, however due to the nature of VGI, thematic attribute completeness needs to be examined separately, as corresponding features may exist, but with no thematic attributes (Maué and Schade, 2008). The feature completeness needs to refer to omission (missing data) and commission (excess data). For linear datasets, measuring the features' length that is matched during the matching process could answer the data completeness question.

For attribute accuracy, the provided approaches usually refer to standardised data, where errors are usually considered as a case of misclassification (Caprioli *et al.*, 2003; Servigne *et al.*, 2006; Devillers and Jeansoulin, 2006), which is not applicable in case of missing attributes, undefined range of values, misspelling or use of abbreviations (Al-Bakri & Fairbairn, 2010). Attribute accuracy for VGI needs to compare the attributes of corresponding features, which links back to the data matching procedure. The provided algorithms for text comparison are 'Soundex', 'Metaphone', 'Levenshtein' and 'similar_text'. The last one is considered as most appropriate: it does not rely on pronunciation, which would restrict this research to a specific area and language; it deals with the use of abbreviations without the need of a separate abbreviations index file, which again would restrict the research scope; it is computationally less complex and, as a result, faster.

For positional accuracy of linear datasets, the different ways of data capture and representation in VGI does not allow an approach based on the position of end-points, well-defined or other sample points (Goodchild and Hunter, 1997; van Niel and McVicar, 2002). Evaluation of linear data by using the epsilon band (Chrishman, 1989; Shi, 1998) is not appropriate for VGI, because it is very sensitive to outliers (Goodchild and Hunter, 1997). The use of buffers as proposed by Tveite and Langaas (1999) is more appropriate for cases where no dataset is considered of higher quality. The Increasing Buffer Method (IBM) (Goodchild and Hunter, 1997) seems more appropriate for the case of VGI, because it is applied on a reference dataset considered as of higher accuracy, it is relatively insensitive to outliers, it is statistically based and does not require matching points between the datasets.

Chapter 4

Methodology

4. Methodology[4]

4.1. Introduction

Following the literature review on VGI, spatial quality and relevant research on VGI quality, this chapter highlights the gaps in the literature that this thesis aims to cover and presents the research objectives. It further continues by describing the suggested methodology that forms the framework for the quality evaluation of VGI linear datasets, and concludes describing the datasets that will be used for the case studies.

4.2. Gaps in the Literature

Although there are already several studies on VGI spatial data quality, as reviewed in section 3.4, there are aspects not covered so far. Specifically, when quality evaluation implies a comparison of a VGI linear dataset with a reference dataset, some of the gaps found in the literature are as follows:

- Existing manual data matching or other manual data corrections, which may be necessary to enable the measurement of quality elements (Ather, 2009; Basiouka, 2009; Kounadi, 2009; Cipeluch *et al.*, 2010; Girres and Touya, 2010; Ueberschlag, 2010), hinder the replication of the method in a different or larger area, or when VGI is updated. Considering the frequency of updates in VGI, this may quickly render quality results obsolete. There is a need to automate the data matching procedure, which requires an approach similar to the ones followed in the different research area of conflation or data fusion, yet customised for VGI.
- Automated data matching that relies solely on geometric attributes (Gabay and Doytsher, 2000; Mustière and Devogele, 2008) is not as effective as when also relying on thematic attributes. However, the matching algorithm must take into consideration that VGI may not have attributes at all, or they may be misspelled. An effective automated data matching

[4] Sections 4.4, 4.6, 4.7, 4.8, 4.9 and 4.10 have been partially adapted from:
Koukoletsos, T., Haklay, M. and Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS* [in press - DOI: 10.1111/j.1467-9671.2012.01304.x].
Sections 4.8 and 4.12 have been partially adapted from:
Koukoletsos, T., Haklay, M. And Ellul, C., 2011. An automated method to assess Data Completeness and Positional Accuracy of OpenStreetMap. Presented at *The 11th International Conference on GeoComputation*, London, UK, 20-22 Jul 2011.

procedure is essential for the further quality evaluation, as it ensures that corresponding objects are compared.

- The matching approach needs to be customised for heterogeneous datasets, where some data or their attributes may be missing from one dataset. Additionally, the provided information that could generally be compared, regardless of the data source, needs to be taken into consideration. In this way the quality evaluation method avoids being suitable for a specific dataset or purpose (e.g. Ludwig *et al.*, 2010), but instead could be applied in any VGI case.
- Lack of automation is also the reason for examining a sample of data within the area studied, usually manually selected (Al-Bakri and Fairbairn, 2010; Girres and Touya, 2010; Ueberschlag, 2010). No matter how representative of the whole dataset this selection may seem, results still refer to sample data and not to the whole dataset for the area examined.
- Indirectly dealing with data matching by comparing selected road types (Ather, 2009; Cipeluch *et al.*, 2010; Haklay, 2010c) may cause data rejection from the evaluation process (in cases of erroneous or no attributes), giving a rather false completeness assessment.
- Quality elements need to be measured using appropriate indicators. Some examples: Data completeness that relies on dataset length with no previous feature matching (Zielstra and Zipf, 2010) does not take into account that dataset 'A' may include additional information not present in the dataset 'B'. Positional accuracy should be described by a distance value, different for each area studied, rather than by a varying overlap percentage for a pre-defined distance (Kounadi, 2009; Ather, 2009; Basiouka, 2009; Ueberschlag, 2010; Haklay, 2010c). Attribute completeness that relies on the number of distinct attribute values (Kounadi, 2009; Ueberschlag, 2010) does not take into account partially missing or misspelled information, failing to notice in the first and optimistically miscalculating in the second case.

4.3. Research Objectives

Section 1.9 presented the general aims and questions of this thesis, while the literature review of Chapters 2 and 3 helped identify the above mentioned gaps in the literature. As a result, the following research objectives are formed and tackled in this thesis.

- **Understand the nature of VGI linear data:** The first objective is to understand the nature of VGI linear data. This includes finding out the general characteristics of a VGI dataset that are common regardless of the source, so they can be used in a comparison procedure. It is

essential to define the general parameters of comparison, such as what is the basic object unit to be compared, what is the significance of the spatial or non-spatial attributes, how to perform the analysis in order to deal with VGI heterogeneity, what quality elements will be measured and what results would be meaningful for those interested in using VGI.

- **Develop a suitable automated data matching procedure:** It is important that any quality analysis based on comparison is performed on corresponding objects. The second objective is to develop an appropriate data matching algorithm, taking into account VGI characteristics from the first Objective. The significance of spatial and non-spatial attributes will define the order of examining them. Lack of information should also be considered when designing the method, so as to use different approaches of finding the corresponding objects.
- **Perform quality analysis:** After having decided on the quality elements in the first objective, appropriate indicators need to be selected and measured. The efficiency of the data matching process of the second objective should be considered, so as to minimise the influence of data matching errors on the quality assessment. The first objective also defines the boundaries for the quality analysis; VGI heterogeneity needs to be considered and VGI frequency of updates demands for an automated procedure to enable future comparison, so the quality analysis process needs to be integrated with the data matching process of the second objective, and together they need to form an automated quality analysis framework. Spatial correlation of quality results is also worth examining, e.g. if data are complete in an area, does it mean that they are also accurate?

4.4. Methodology overview

To reach these objectives, a comparison process is designed as presented in the flow diagram of Figure 4.1. After selecting VGI and reference sources, some data preparation is necessary to load the datasets in a PostGIS database (see next section). A tessellation file covering the study area is also necessary to deal with VGI heterogeneity, as well as for computational reasons (see sections 4.6 and 4.15.1). The whole process is developed as a web-page application to enable user interaction when necessary, using PHP (see Appendix 'A'). The process begins by selecting one by one the tiles of the tessellation file. Each tile is used to clip both datasets (section 4.6) and the analysis is performed for each tile individually. The tile is classified according to the network density as 'urban' or 'rural' (section 4.7). The data matching process, consisting of seven successive stages, begins (section 4.8) and leads to two subsets for each dataset, containing data present in both datasets and data unique in each dataset. The process then moves to evaluate data quality elements. Data completeness (section 4.10) relies on data matching and uses the subsets created. Attribute accuracy follows,

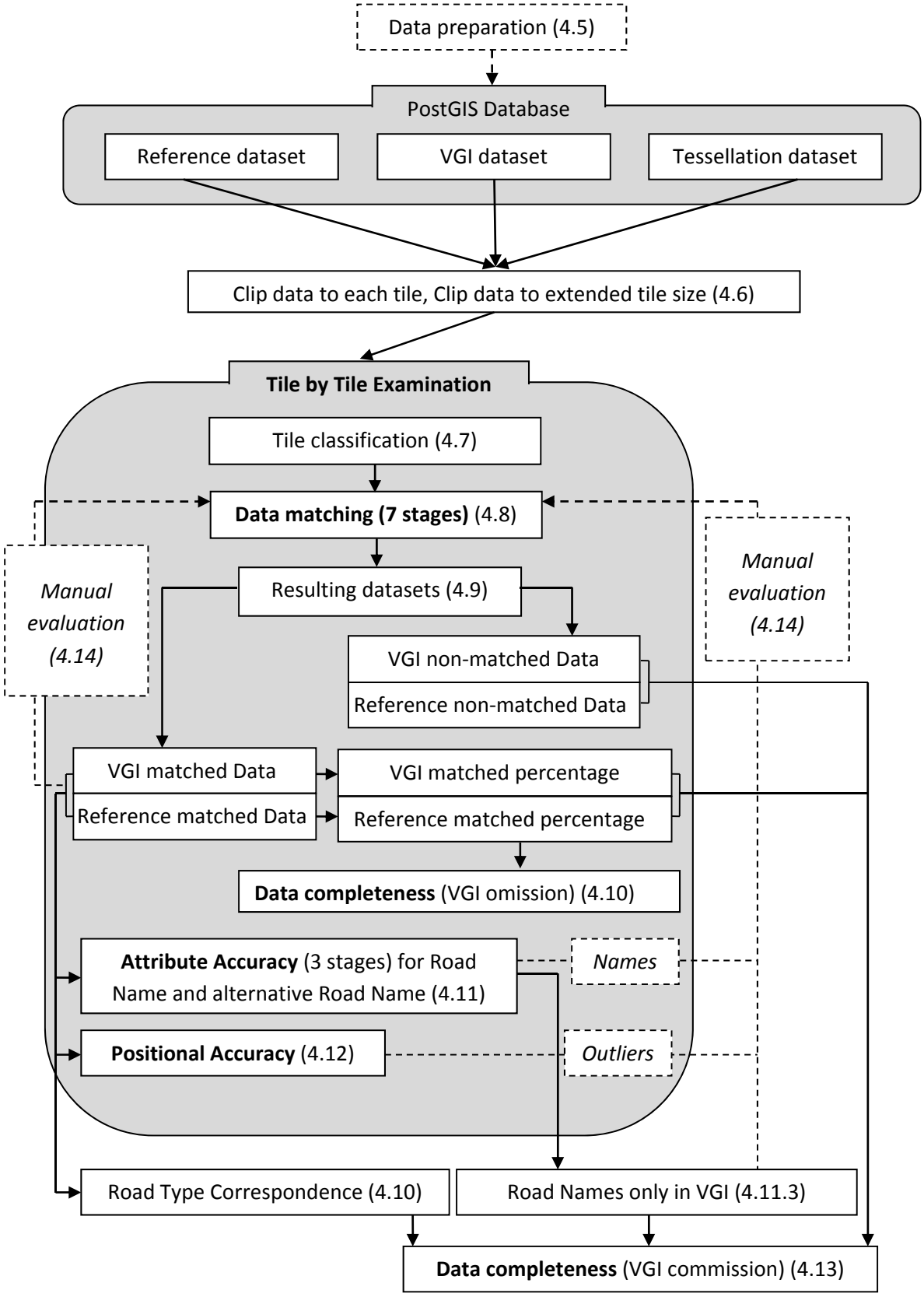

Figure 4.1: *Flow diagram of the developed methodology (with section index in parenthesis). Data matching (7 stages) is further analysed in Figure 4.2.*

including 3 stages and using the matched subsets (section 4.11). Positional accuracy, finally, uses the IBM mentioned in section 3.3.4 to apply varying buffers on the reference matched subset in order to measure the distance of the VGI matched subset (section 4.12). Tile examination ends here and quality results for the tile are stored before moving on to the next tile. Information regarding the whole dataset is also stored and will be combined with results from other tiles in the end (e.g. road types' correspondence or indication of VGI commissioned data) (section 4.13).

The parameters that will be used in this approach are gathered in Tables 4.7 and 4.8 (section 4.15), with links to the sections where they are discussed and justified. This was considered necessary because justifying some of them during the description of the successive computational steps might have confused the reader, by steering away from the conceptual framework. Thus, some of the parameters are explained and justified separately or even after presenting the case studies.

4.5. Data preparation

Once the vector datasets are acquired, they need to be in the same coordinate system, so re-projection of one or both datasets maybe necessary. The tessellation file is also necessary to be in the same coordinate system. Datasets then need to be uploaded in a PostGIS database. When running the process, the application enables the user to login to the database using the appropriate credentials and select the tessellation file and datasets to compare (see Appendix 'A' for more details on the application interface). More information on data preparation is given in each case study individually (see sections 5.2, 6.2, 7.2).

4.6. Reference and VGI clipping

As mentioned in the previous chapters, one way of dealing with VGI heterogeneity is to localise the evaluation and produce results for discrete areas. This can be achieved using a tessellation file, which is a collection of adjacent polygons that can be used to clip the datasets. As a result, each cell is processed individually and local measurements of data quality are possible. Additionally, computation is faster by limiting the number of objects processed each time. There are, however, two issues of tile size and tile shape, which are further discussed in section 4.15.1.

Regardless of the tile size or shape, corresponding objects that are parallel and close to the tile borders may lie in different tiles and will fail to be examined and matched. Additionally, erroneous matching can occur for clipped objects when only a small part of them is examined, because some of

the geometric constraints used in data matching rely on features' length (see section 4.8.3). To solve this, an extended tile is created by buffering the initial tile by 50 m. The selection of this buffer width is justified in sections 4.8.3 and 8.2.2. The extended tile is also used to clip the datasets, so actually there are two pairs of sub-datasets, the ones referring to the core tile size (e.g. 1 km^2) and the ones referring to the extended one, covering a slightly larger area (e.g. 1.21 km^2). The matching algorithm is applied on the extended sub-datasets and the erroneous data matching moves to the extended tile borders. When the tile examination finishes, the resulting datasets (matched and non-matched data) are clipped to the initial tile size, so the erroneous data matching that may occur next to the extended tile border is removed. More details on the benefits of using the extended tile are presented in section 8.2.2.

4.7. Tile classification

After experimenting with the datasets, it was found that in rural areas some corresponding objects may be found in far larger distances than in urban areas. This can be attributed to three factors: reduced satellite imagery resolution that can lead to mislocated objects by hundreds of meters in VGI (Ramm *et al.*, 2011, p.136), reduced official data accuracy (e.g. Ordnance Survey, 2009c) or other unknown user-specific reasons. As a result, data matching can be more efficient by using looser constraints in rural and stricter ones in urban areas, where the road network is denser and the accuracy is anticipated to be better. This requirement is also mentioned by Walter and Fritsch (2001).

However, there is usually no simple way to classify rural or urban areas. In two of the case studies, the reference dataset provides no metadata on where urban switches to rural or moorland accuracy. Likewise, VGI rarely provides information on the source of data: within the same area some features may have been mapped using a GPS receiver, while some others are the result of digitising satellite imagery. Additionally, in several cases rural and urban boundaries may be fuzzy; rural areas may contain smaller areas of increased network density, which should be addressed as urban, and likewise the suburbs of a city may as well have a density similar to a rural area and no high resolution satellite image coverage.

In this study, tile classification takes into consideration the road density by counting entities and junctions. After tests, it was decided to classify a tile as rural when both reference and VGI datasets contain less than 17 entities and 8 junctions per km^2. This selection is further justified in section 8.2.3.

4.8. Data matching (7 stages)

4.8.1. Data matching overview

The data matching procedure, briefly mentioned in the flow diagram (Figure 4.1), can be further analysed into its seven stages, as shown in Figure 4.2. This paragraph is a brief description of the data matching process. Features of both datasets are divided into segments (the terms 'feature' and 'segment' were explained in section 3.3) and data matching begins. The first four stages deal with segments and the last 3 are at the feature level. The first level relies on geometry and deals with '1-1' segment matching. Levels 2 and 3 add an attribute constraint, looking for exact and similar road names respectively. Level 4 deals with segments with no name attribute. Level 5 recomposes the features and classifies them as matched or not. Levels 6 and 7 examine VGI features with and without road names respectively. Finally, matched and non-matched data subsets are produced for both datasets.

Data matching (7 stages)

Feature segmentation (4.8.2)

Ordered stages

Stage	Basic Unit	Constraints (in order of importance)	Section
1	Reference Segment	Geometric (Distance,Orientation,Length)	4.8.3
2	Reference Segment	Attribute and geometric (Name, Distance,Orientation)	4.8.4
3	Reference Segment	Attribute and geometric (Name, Distance,Orientation)	4.8.5
4	Reference Segment	Geometric (Distance,Orientation)	4.8.6
5	VGI and Reference Feature	Geometric (Length)	4.8.7
6	VGI Feature	Attribute and geometric (Name, Distance)	4.8.8
7	VGI Feature	Geometric and attribute (Distance, Length)	4.8.9

Figure 4.2: *Flow diagram of the data matching process*

4.8.2. Feature segmentation

The basic object unit for both datasets is the feature, ranging from a small straight line to a complex polyline with many vertices that represents a road (partially or as a whole). A feature in one dataset

may correspond to more than one in the other, which makes automatic matching of features difficult. Therefore, features of both datasets are divided into segments (see section 3.3 for term definitions). Segments' length, direction and position depend on data capture or digitisation. Dealing with segments enables directional analysis, which is problematic if having to deal with features (De Smith *et al.*, 2009). The first four stages of the matching algorithm deal with segments. For each segment a new identification number (ID) is assigned to enable segment-by-segment processing, while their original ID is retained for the re-composition and examination of the features in the final three stages.

Length (***S***) is calculated for each segment, using Equation 1. If zero-length segments are created during the segmentation, they are removed. Successively, orientation ($\boldsymbol{\vartheta}$) is calculated for each segment, using Equation 2. These parameters are necessary for the geometric constraints.

$$S = \sqrt{(x_{end} - x_{start})^2 + (y_{end} - y_{start})^2} \quad (1)$$

$$\vartheta = \frac{180}{\pi} \times \text{atan2}\ ((x_{end} - x_{start}),\ (y_{end} - y_{start})) \quad (2)$$

In Equation 2, 'atan2' function returns radians, which are transformed to degrees when multiplied by 180/π. Direction is not important in our analysis, so all angular results are further converted to a range from 0 to 180 degrees, which renders their comparison easier.

4.8.3. Stage 1

The matching process starts by testing the reference dataset segment by segment to find the best VGI matching candidate. The reason for this is because the reference dataset is considered to be complete and additionally to have better topology, especially when built for routing purposes. This means that intersected roads will have a road junction if necessary, so road junctions are expected to be at one end of a feature (or its segmented part). This does not necessarily happen in VGI, and moreover VGI may contain additional information not collected by the reference dataset; examining each VGI segment instead of each reference one would result in trying to match non corresponding objects (e.g. VGI's steps or footpaths with something completely different from the reference dataset). The topology issue is further discussed in each case study chapter (sections 5.5.4, 6.5.4, 7.5.4).

For each reference segment, possible corresponding VGI segments are sought, using the following geometric constrains to narrow down the results:

Search distance: Considering that a reference segment represents the true position of a road (or part of it), it is assumed that VGI mapping varies according to the GPS receiver accuracy. Ramm *et al.* (2011) suggest an average GPS accuracy of 5 m for VGI mappers, however a more conservative approach is followed here. The search distance (**D**) is defined by the Equation 3:

$$\mathrm{D} = \mathrm{c} \times \mathrm{a} + \frac{\mathrm{w}}{2} \tag{3}$$

where**:**

- **'a'** is the assumed GPS accuracy, considered as 10 and 15 meters for urban and rural areas respectively. In reality GPS is not expected to be less accurate in rural areas, however the latter value is used to enable looser rural constraints, by increasing the search distance and 'angular tolerance', explained later on.
- '**c'** is an integer (2 for urban and 3 for rural areas), used to cover worst case scenarios such as lower quality of GPS receivers, multipath rejection or bad signal reception when mapping urban canyons, cases where dual carriageway motorways are represented as a single line in one dataset, reduced satellite imagery accuracy in rural areas, and digitisation errors in VGI.
- **'w'** is an assumption of the reference road width based on the road type and can be adjusted to other reference datasets characteristics. By adding half of it, the search distance is increased to cover cases where the VGI mapper moves along the side of the road (e.g. pavement) instead of its axis.

Assuming road widths from 2 m (Alleys) to 11 m (Motorways) for the equation (3), the search distance for stage 1 will be 21 to 26 meters in urban and 46 to 51 meters in rural areas. Further discussion and justification of the selected values for a, c and w is provided in section 4.15.2.

Orientation: For the VGI segments found within the above distance, orientation is examined. An angular tolerance (***φ***) is needed to identify segments with orientation similar to the one examined. In Figure 4.3a, reference segment of length **β** is represented by a continuous line, while the dotted lines represent some possible scenarios of mapping it with GPS, with its average accuracy represented as a circle of radius **α**. Figure 4.3b shows the worst case scenario, leading to the Equation 4.

$$\varphi = \frac{180}{\pi} \times \arcsin\left(\frac{\alpha}{\beta/2}\right), \quad \text{computationally simplified to} \quad \varphi = \frac{180}{\pi} \times \arctan\left(\frac{\alpha}{\beta/2}\right) \qquad (4)$$

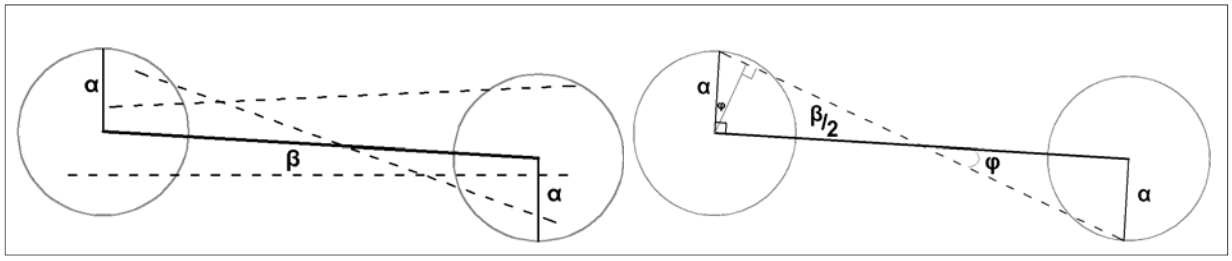

Figure 4.3a: *Possible scenarios and* ***b:*** *worst case scenario for the calculation of angular tolerance*

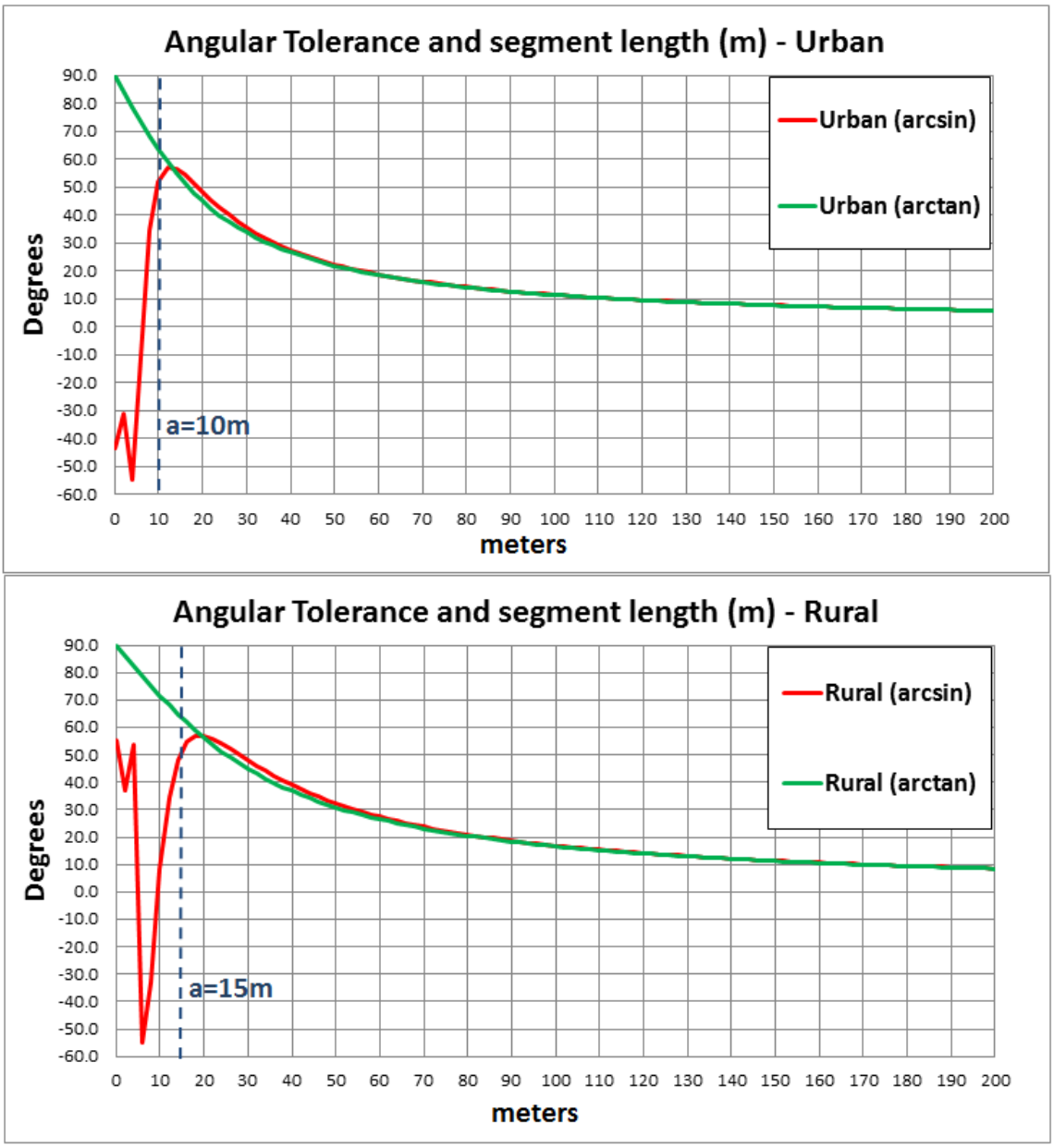

Figure 4.4: *Results from actual (red) and simplified (green) equation for angular tolerance*

Figure 4.4 justifies the use of the simplified equation (4) (green line). For segment lengths smaller than the assumed GPS accuracy 'a', the simplified equation does not require distinctive computational analysis, which would slow down the matching procedure significantly. For longer segments (up to 60 m) the simplified version produces slightly smaller angular tolerances, so segments with larger variations in orientation that should be considered may be rejected (false negative or type II error), which is a more conservative matching approach. For even longer segments, results remain the same for both the simplified and actual equations.

Figure 4.4 also explains how directional constraints can be looser in rural areas (bigger angular tolerances) by assuming lower GPS accuracy when using equation (4). It further justifies the 50 m buffer while clipping the tiles (mentioned in section 4.6 and further explained in section 8.2.2). Finally, it shows that the smaller the segment is, the bigger the angular tolerance φ will be, and vice versa. Bigger tolerances for smaller segments are needed for matching roundabouts or when matching a feature that is more detailed in one dataset and generalised in the other (Figure 4.5).

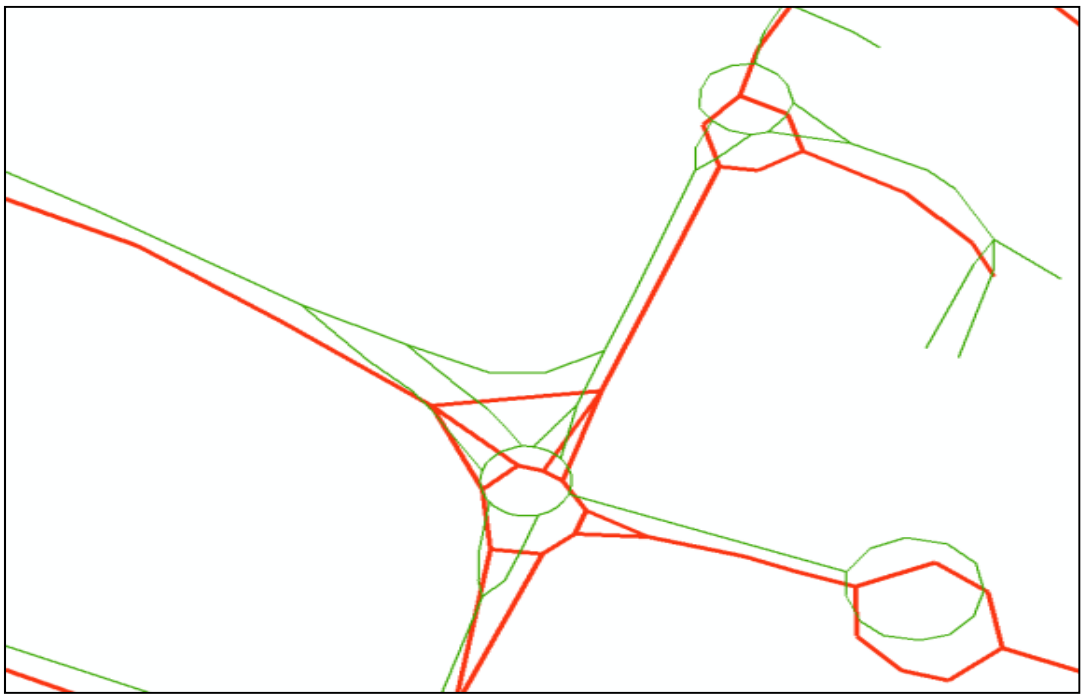

Figure 4.5: *Directional segment matching: Smaller segments demand bigger angular tolerances*

Length: Among the VGI segments found within the above search distance and angular tolerance, only those with length less than three times the reference segment length are considered, in order to avoid matching objects of unusually different length.

These constraints reduce the number of possible matches for each reference segment. At this stage, however, only cases found with one possible match are accepted ('1-1' match), while all the next stages deal with cases of more than one matching candidate. These reference segments are marked, along with the VGI segments that match them, and will not be examined further.

4.8.4. Stage 2

This stage relies mostly on road name attribute. For every non-matched reference segment with a road name attribute, corresponding VGI segments are searched using the same 'Distance' and 'Orientation' constraints as in stage 1. For each VGI candidate found, an exact road name matching between the two datasets is sought. An option to examine a secondary road name attribute is provided, as this can apply to motorways or highways in rural areas with a different conventional naming (i.e. M25, A24). The second name attribute is examined separately, since some roads may have values for both attributes.

4.8.5. Stage 3

An exact name matching between two datasets is not always possible due to misspelling or use of abbreviations in one or both datasets. This stage covers such cases by checking for text similarity. It is applied to each remainder (non-matched) reference segment with a road name attribute, using the same 'Distance' and 'Orientation' constraints as before.

Text similarity constraint: For each VGI segment compared with the reference segment under examination, the number of similar characters is calculated and expressed in percentage (Equation 5).

$$\text{namematch} = \frac{\text{number of similar characters}}{\text{maximum (Reference string length, VGI string length)}} \% \quad (5)$$

Only VGI segments that have a score above 65% are chosen as possible candidates, a threshold calculated through empirical study of the London area. Table 4.1 provides some examples. 65% is quite a low value for a threshold. The limitation of the text similarity approach is that shorter road names are more significantly affected if one character is different (see for example cases 3 and 5 of Table 4.1, where the same abbreviation is used but resulting percentages differ), so a lower threshold would cover most of these cases. On the other hand, however, such a low threshold may regard as similar two completely different road names. For example, 'Abercorn Road' and 'Aberdare

Road' are two completely different roads in the London area that have 69.23% text similarity. Luckily, they are too far from each other to be erroneously matched. To generally avoid such cases, however, among the VGI segments that have a score above 65%, only the ones with the highest score are marked as matched, along with the reference segment under examination. In case that even then two non-corresponding segments are erroneously matched, the error will hopefully be corrected in stage 5.

Index	Reference Road Name	VGI Road Name	'namematch'
1	ST ANNE STREET	St. Annes St	68.75 %
2	HALSEY STREET	Halsey St.	69.23 %
3	ST MALO AVENUE	St Malo Ave	78.57 %
4	ST MARY'S MEWS	St Maryâ€™s Mews	81.25 %
5	NORTHUMBERLAND AVENUE	Northumberland Ave	85.70 %
6	QUEENS GATE	Queen's Gate	91.67 %
7	ST JOHN'S HILL	St. John's Hill	93.33 %

Table 4.1: *String similarity scores for various cases*

Due to the nature of the secondary name attribute (further discussed in section 4.11.3), text similarity is only applied on the primary road name attribute.

4.8.6. Stage 4

This stage deals with reference segments not matched so far, also targeting those with no name attribute, so it will be based solely on geometry. For each such reference segment, the same 'Distance' and 'Orientation' constraints are used as in previous stages. For every VGI segment found, the distances between reference and VGI start-points and end-points are calculated correspondingly. The sum of these values becomes an attribute for each VGI candidate, and the segment with the minimum value is selected.

Figure 4.6: *Data matching challenges in stage 4*

This approach was proven to give better results than checking the distance between linear segments, or between their centers. Figure 4.6 provides an example: reference segment R1 will be matched with VGI segment V1, despite that V2 is closer to it.

4.8.7. Stage 5

The previous stage was the final one dealing with segments. In this stage the features are recomposed and are classified as having a corresponding one in the other dataset or not, using the segment-by-segment matching information.

Figure 4.7a: *Datasets before the matching process,* ***b:*** *Matched segments before stage 5,* ***c:*** *Matched features after stage 5*

For each reference and VGI feature, the sum of segments' length found as matched is compared to the total feature length. If more than half of the feature length is already found as matched, then this feature is considered to have a match. In this way, this stage can deal with possible errors in previous stages due to automation. An example can be seen in Figure 4.7, where (a) represents the reference (green) and VGI (red) datasets under comparison, (b) shows the matched segments before stage 5 and (c) shows the matched features after stage 5. In Figure 4.7b there are missing segments (failed to be matched) to the North, while there are also segments mistakenly matched in the centre and South. In Figure 4.7c stage 5 succeeds in correcting these data matching errors, matching the features quite efficiently.

Using the segments' correspondence, a feature correspondence is also created. The feature's segment (or segments) with the longest length that was (or were) linked to segments belonging to one and the same feature on the other dataset, will help create a link between the two features. This new attribute will be used for attribute accuracy assessment (section 4.11). For the feature data matching, however, which will also be used for data completeness assessment (section 4.10), the classification of a feature, as matched or non-matched is enough.

4.8.8. Stage 6

Stage 5 may also lead to errors when non-matched segments from previous stages are bigger than the matched ones, as a result of the conservative approach, leading to mistakenly classifying features as non-matched ones. Additionally, for the first 4 levels the reference data are examined as compared to the VGI dataset, which may skip some VGI objects. This and the next stage compares VGI objects against the reference dataset. In this stage each VGI non-matched feature with a name attribute is examined. For each one, reference features with a similar name are searched within a distance, which is c times the GPS assumed accuracy (as described in Stage 1). This results in 20 and 45 meters for urban and rural areas respectively, if using the values described in section 4.8.3. Name matching follows the text similarity approach of stage 3, however with a threshold percentage raised to 75%.

Section 4.8.5 discussed the use of 65% as a text similarity threshold, mentioning that since it is a low threshold, different road names may also be considered as similar. This was partially solved by matching only the segments with the highest percentage. When dealing with features instead of segments, however, the number of candidates (if any) will be lower, usually one. Additionally, this stage accepts all features above the threshold and not only those with the highest percentage, so as

to cover cases of different representation (e.g. one feature corresponding to two in the other dataset). Raising the threshold to 75% keeps the approach quite conservative, preffering to avoid matching non-corresponding features than missing some that should have been matched. This value was also selected by testing road name matching in London area. This is further discussed in section 8.5.

Concluding, apart from corresponding features that lie next to each other (which for some reason have not already been matched by the previous stages), succesive features with the same attribute name within the search area are also covered, even if one of them is mispelled or has an abbreviation (as long as the score is above 75%). The same process is applied on the alternative name attribute, this time using exact name matching instead of text similarity.

4.8.9. Stage 7

This stage deals with VGI non-matched features with no name attribute. Additionally, tests showed that the procedure so far may fail to find a match for some features of parallel VGI roads that are close to each other, such as the cases of motorways with many carriageways, especially near junctions and slip roads. Due to their distance or naming, the matching algorithm assigns the adjacent feature as best candidate in previous stages, leaving a gap in the bigger object represented in the VGI dataset (Figure 4.8).

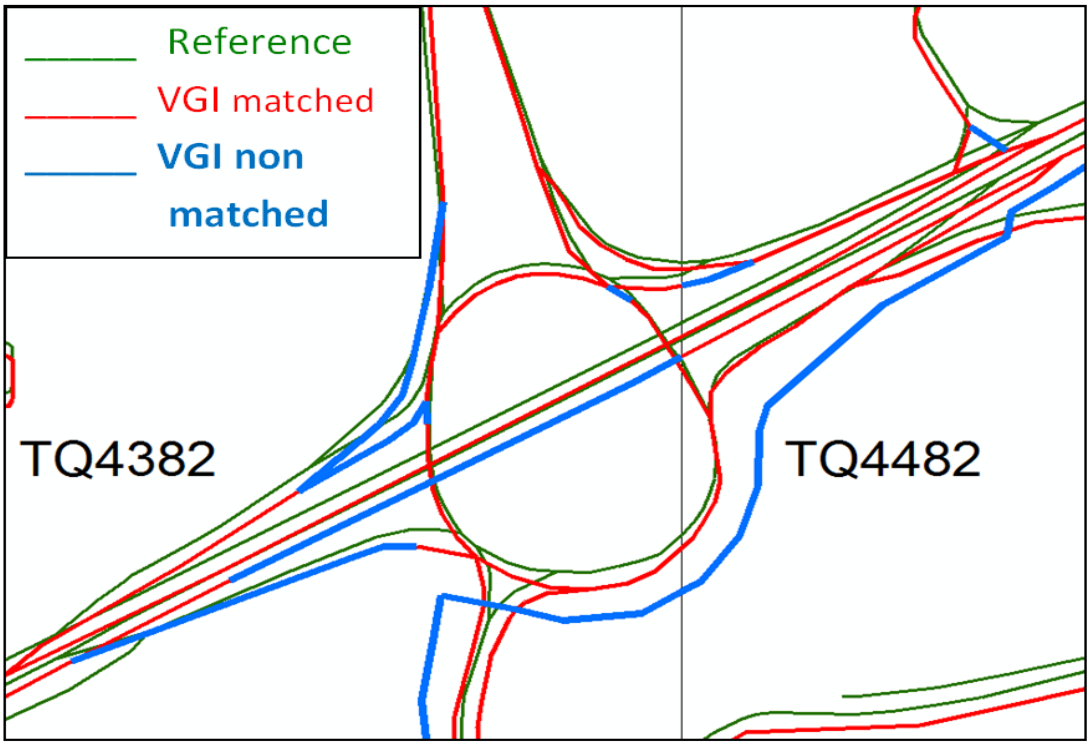

Figure 4.8: *Matching errors before stage 7*

For each non-matched VGI feature, a buffer is applied of a width that equals to the GPS accuracy. The reference features' length inside this buffer is calculated, as well as the VGI matched-so-far features' length. These lengths are compared with the length of the feature under examination and Equations 6 and 7 are used to decide whether the VGI feature has a match.

$$\text{Ref. length inside buffer} - 2 \times \text{GPS} > 0.8 \times \text{VGI non-matched feature length} \quad (6)$$

$$\text{VGI length inside buffer} - 2 \times \text{GPS} < 0.9 \times \text{VGI non-matched feature length} \quad (7)$$

The GPS accuracy value is subtracted twice, since buffer area extends all around the feature. The 0.8 and 0.9 factors are used to cover cases of simpler (generalised) lines with slightly lower length values. Combining the above restrictions, a VGI non-matched feature should have a bigger length than the matched-so-far VGI features and smaller length than the reference features inside the buffer to be considered as a match.

The 0.8 and 0.9 factors were the result of testing different values in three cases similar to the one presented in Figure 4.8. Equation 7 succeeds in avoiding matching VGI parallel features that do not comply with stage 4 by not being the closest ones. As an example, a VGI cycleway or footpath with no road name information that runs parallel to an already matched VGI feature representing a road will not be matched in this stage.

Depending on the datasets involved in a comparison, the level of generalisation between them may differ, for example if one dataset is more generalised, lower values than 0.8 and 0.9 may need to be used. However, the efficiency of this stage in this case is the least of the problems, as generalisation may affect other stages as well (e.g. search distance cannot reach the distance of corresponding objects, or their orientation may differ more than the angular tolerance). This is mentioned as a limitation of the data matching method in section 8.8. Although these values seem appropriate for the case studies and the different datasets used in this thesis, they need to be further examined through a sensitivity analysis. Their decision is based on a trial-and-error basis: after manual evaluation of the data matching process of all case studies, data matching errors were not specifically created during this stage. Due to the low data matching errors, on the other hand, further optimisation of these values was not considered.

4.9. Resulting datasets

Section 4.6 mentioned the use of the extended tile during the matching procedure. The resulting extended sub-datasets now contain an additional attribute to describe if they were matched with a feature in the other dataset or not. They are clipped using the initial tile, so that possible matching errors next to the extended tile border are rejected. For each dataset, features marked as matched are collected into a new sub-dataset, called 'matched', while the rest create the 'non-matched' dataset. Matched sub-datasets contain data that represent objects present in both datasets, while non-matched datasets contain data unique for each dataset. The first subsets, that contain only corresponding data, allow for a more meaningful further implementation of data quality tools, while the second ones provide information on the excess data (over-completeness or commission) and can be useful for conflation purposes.

4.10. Data completeness (VGI omission and commission)

Quality evaluation begins with data completeness. The total length of matched features is calculated and compared with the total length of each dataset for each tile. The matched length, expressed in percentage, shows the amount of data found in the other dataset. Specifically, the reference dataset's matching percentage is the percentage of reference data also found in the VGI dataset, so this represents VGI completeness (as compared to the reference dataset). On the other hand, VGI matching percentage is the percentage of VGI data also found in the reference dataset. A low value means that there are features with no correspondence in the other dataset, resulting in a dataset with additional information. The average of the two percentages for each tile could be used as a mixed percentage to show the level of agreement between the datasets. Four cases can be roughly distinguished (Table 4.2). Figure 4.9 provides an example (derived from the first case study).

Case	VGI matching percentage	Reference matching percentage	Mixed percentage	Meaning
1	High	High	High	Datasets agree with each other
2	High	Low	Low	Reference dataset is richer
3	Low	High	Low	VGI dataset is richer
4	Low	Low	Very Low	Datasets contain different data

Table 4.2: *General cases of matching for each cell*

Figure 4.9a: *Reference & VGI datasets,* ***b:*** *Reference matching percentages (VGI completeness),* ***c:*** *VGI matching percentages,* ***d:*** *Mixed matching percentages*

The classification used in all similar figures in this thesis needs to be further explained. Using many classes gives a better insight in data distribution. Seven classes are used here, as this is the maximum number of classes suggested for visualisation purposes (Kraak and Ormeling, 1996). Although quality results may significantly differ between the two datasets, the same classification needs to be followed in each quality element measurement so that this difference is appropriately visualised. The next issue is the range of each class. Although there are various ways of classification (e.g. natural breaks, quantiles, equal or geometrical intervals, etc), classes need to have reasonable boundaries of integer values, while at the same time distribute the values as evenly as possible. This, however, cannot be effectively applied due to the different nature of datasets even within the same area. A manual classification will be selected in each case study, customised to combine all the above (if possible). It is considered important to have a separate class for zero values, which is indicative of information not present in the other dataset or totally inconsistent data and can narrow down post-processing evaluation. Similarly, a separate class for the value of 100% shows total agreement of one dataset compared to the other, which also narrows down post-processing evaluation. The second to fourth classes are evenly distributed from 0 to 75%, while fourth and fifth

classes split the remaining range in 75% to 90% and 90 to 99.99%. This was decided because of the good quality of data in all case studies: as most tiles score above 75%, there is a need for a further division to 'good' and 'better' quality. Other data sources may require different classification.

While VGI completeness (omission of data) can be obtained from the reference data matching percentages, VGI commission of data (existence of additional data) requires a more complicated judgement, because VGI data matching percentage refers to all VGI objects, including information that may not be represented in the reference dataset (e.g. footpaths). Information from the VGI dataset that is not included in the specifications of the reference dataset could be rejected by excluding selected VGI road types before the analysis (e.g. footpaths). However, it is not certain that VGI users classify road types correctly, so valuable information might also be rejected.

Reference type	VGI type	Matched %	VGI type	Reference type	Matched %
B Road	secondary	91.2	**secondary**	B Road	86.6
	residential	3.1		Minor Road	5.9
	unclassified	2.2		Local Street	3
	primary	1.1		A Road	2.9
	tertiary	0.8		Private Road - Restricted Access	1.4
	trunk	0.5		Alley	0.1
	trunk_link	0.4		Private Road - Publicly Accessible	0.1
	service	0.2		Pedestrianised Street	0
	footway	0.2			
	cycleway	0.1			
	steps	0			
	pedestrian	0			
	path	0			
	road	0			
	bridleway	0			
	primary_link	0			
	private	0			

Table 4.3: *Reference road type correspondence*

To avoid rejecting corresponding data, which would give a wrong impression of VGI completeness, all VGI information is used. Based on the object correspondence during the matching process (stage 5), information about road types' correspondence is also collected. In a separate table the length between corresponding road types is added (e.g. 352 m of 'A Road' correspond to 331 m of

'Primary Road'). When tile by tile examination finishes and all possible road type combinations are collected for the whole dataset, a percentage is calculated as the length of each pair compared to the total length of the examined road type. This is applied in both datasets. Table 4.3 provides an example for reference 'B Road' and VGI 'secondary road', which bilaterally prove to be corresponding road types (results from first case study).

This information is used to move deeper into VGI commission, which is also a part of data completeness. More details are provided in section 4.13.2.

4.11. Attribute accuracy

4.11.1. General

Due to the different structure and objective of the datasets under comparison, attribute information will generally be different. For example, an official dataset may contain routing attributes, while VGI may not if it is not created for routing purposes. In order to assess attribute accuracy, common attributes need to be found. The most common one in road network datasets is the road name. Road type is also an attribute present in most datasets, but usually the different classification used in each data source for the same network hinders the comparison. However, both these are attributes that can generally be found in any VGI and official data source. The approach followed in this study compares the road name and allows also for a second name attribute comparison, in case of alternative or supplementary name. Information is also collected on the road types regarding their correspondence between the datasets, as section 4.10 mentioned. Using only name attributes broadens the scope of the analysis to other linear datasets as well (e.g. water networks).

A possible approach for evaluating attribute accuracy would be to examine the distinct names within the area examined (tile). However, simply by counting and comparing the total number of road names for both datasets would not always be a useful indicator because of the nature of VGI. Misspelling or use of abbreviations in some of the features that form a road increases the number of distinct values, giving a false impression of a richer VGI dataset. If at the same time some attribute information is missing in other features, the number of distinct names may not be higher than the one of the reference dataset, failing to show both the additional and the missing attribute names. Additionally, when counting the distinct names in a tile, the case of erroneous swapping of values in VGI features is not covered. Road A and B may have mistakenly been named B and A in VGI

respectively, however the number of distinct names remains the same while attributes are not accurate.

Calculating the length of features with correct attributes is a more valuable indicator than distinct names, because it also covers cases where attribute information is partially missing in VGI (e.g. when one of the many features comprising a road is left with no name attribute).

Attribute completeness and accuracy are dealt more efficiently when moving to the feature level by comparing names of the corresponding features, rather than generally comparing all the names within a tile and ignoring their spatial relationship. This can be achieved by using the feature correspondence found in stage 5 of the matching process (see section 4.8.7). However, since automation may lead to data matching errors, attribute accuracy approach should be designed accordingly to avoid them, if possible.

4.11.2. Method description

Section 3.3.3 described the text similarity function as the selected algorithm to deal with the misspelling or use of abbreviations, which leads to slightly different naming for the same object between the datasets. The proposed approach is applied on the subsets of matched data and includes the following three stages.

Stage 1:

This stage assumes that features are matched correctly. For the tile examined, each feature with a name attribute is processed. Its name is compared with the corresponding feature (as found in stage 5 of the matching process) for exact name matching. If matching is found, it is marked accordingly (by being given a value ‘1’) and will not be examined further.

Stage 2:

This stage also assumes that features are matched correctly, but targets cases of misspelling or abbreviations. Each feature with a name attribute that was not marked in the previous stage is compared with the corresponding feature, this time looking for the level of text similarity. If names bear similarities above 70%, the feature is marked accordingly (by being given a value ‘2’) and will not be examined further.

The 70% threshold is decided to be between the lower 65% threshold that was used in stage 3 of the data matching procedure (discussed in section 4.8.5) and the higher 75% that was used in data matching stage 6 (discussed in section 4.8.8). Using road name comparisons in the London area, as already mentioned, the lower threshold allowed some different names to be considered as similar, while the higher and more conservative one rejected similar names as different (mostly cases of short road names). This is further discussed in section 8.5.

Stage 3:

Unlike the previous two stages, this stage handles cases of erroneous feature matching. Each feature with a name attribute that was not marked in the previous stages is processed. The linked feature is ignored. A search distance is defined (35 m for urban and 50 m for rural tiles, as classified in the beginning of the data matching process – section 4.7), within which the reference road names are examined for text similarity and are marked accordingly (being given a value '3') if they are similar over 70%. Search distances are decided to be wide enough for the anticipated accuracy of the area, but at the same time small enough to limit the examined cases and avoid regarding attributes as correct for objects that are not relatively close to each other.

4.11.3. Attribute accuracy assessment

This technique is applied on both the reference and VGI matched subsets for the primary name attribute, while stage 1 is also applied on the secondary one. The reason for not looking for text similarity for the secondary name attribute is that in most cases the nature of the name does not allow it. For example, a second name attribute could be 'M25', 'A240', 'B483'. This conventional road naming is usually followed for Motorways, A and B Roads, mostly used for highways of arterial roads in rural areas. For this attribute, only exact name matching is suitable. The two names (primary and secondary) need to be examined separately, because sometimes a road may have both attributes present.

When examining each dataset, erroneous assessment is likely to happen close to the tile borders for corresponding features that lie on adjacent tiles. This issue is handled with the use of the extended tile, as mentioned in section 4.6. Specifically, for each dataset under examination, data from dataset's 'A' normal tile are examined against data from dataset's B extended one.

Information is collected for each feature with an attribute, regarding its attribute existence in the other dataset. In the end, all matched features with a name are divided into those with an attribute

value between 1 and 3, according to the previous section (accurate ones), and those with no value (inaccurate or non-existing). For the reference dataset, the length of the features with an attribute value is calculated. This provides the length of the network (within the tile) with a name attribute that is found accurate (in terms of attributes) in the VGI dataset. By comparing it to the total reference length with a name attribute for the tile examined, the percentage of reference dataset found with the same or similar attribute name in VGI is calculated. This is actually the VGI dataset's attribute accuracy, similar to the approach of VGI feature completeness of section 4.10. Likewise, by examining the VGI percentages, 100% value means that existing VGI features with names have correct values, however by itself it does not provide information on missing road names. Lower percentages, on the other hand, may mean either that the name attribute in VGI is wrong (therefore not found to be accurate), or that there are features with attributes in VGI but not in the reference dataset. These are further examined as indicators for VGI commission in section 4.13.4.

There are three percentages calculated, for primary, secondary and total road name attribute accuracy respectively (Equations 8-10):

$$\text{att. acc1} = \frac{\text{accurate name1 length}}{\text{total length with name1 att.}}\,\% \tag{8}$$

$$\text{att. acc2} = \frac{\text{accurate name2 length}}{\text{total length with name2 att.}}\% \tag{9}$$

$$\text{att. acc} = \frac{\text{accurate name1 length} + \text{accurate name2 length}}{\text{total length with name1 att.} + \text{total length with name2 att.}}\% \tag{10}$$

The accurate name1 and name2 length refer to the matched data per tile, while the total lengths refer to the whole tile (including non-matched data with name1 and / or name2 attributes).

Although the attribute accuracy percentage refers to the whole tile, the proposed approach enables attribute accuracy assessment down to the feature level, providing information on which feature's name is exactly matched or bears strong similarity with the official name. Additionally, non-matched VGI dataset is checked for non-matched features with attributes. If any found, they are marked and will be used to indicate VGI commission, as explained in section 4.13.3.

Attribute accuracy results for each tile are represented similarly to data matching results (explained in section 4.10 – Figure 4.9).

4.12. Positional Accuracy

4.12.1. General

Positional accuracy is based on Goodchild and Hunter's (1997) Increasing Buffer Method (IBM), described in section 3.3.4. There are two options of applying the IBM method. The simplified one assumes a positional accuracy value for the tested dataset, which is then used as a buffer width on the reference dataset. The length of the tested dataset that falls within the buffer is calculated and compared to its total length. By using different buffers, the corresponding percentages can be calculated. This way is presented and tested by Goodchild and Hunter (1997), and used thereafter by other studies related to the positional accuracy of linear datasets (Ather, 2009; Kounadi, 2009; Al-Bakri and Fairbairn, 2010; Haklay, 2010c; Ueberschlag, 2010). The drawback of this approach is that positional accuracy is assumed and not calculated. Although Goodchild and Hunter's (1997) case studies and results are also based on the use of discrete values as buffers (simplified version), they propose an algorithm for the inverse procedure (p.302), which is the second and more complex version of the IBM. In this case, positional accuracy is calculated as output (instead of being assumed and used as input) through an iterative process, which uses as input the desired overlap percentage. In other words, the buffer calculated is the one that includes a specified percentage of the tested line when applied on the reference dataset.

The second approach seems more suitable for an automated evaluation of positional accuracy, because the overlap percentage can be regarded as the desired level of confidence, which needs to be decided by the user, while the positional accuracy, represented by the buffer width, has to be accurately calculated instead of being assumed using one or more predefined values. The authors strongly suggest that their proposed iterative algorithm should be used for a positional accuracy evaluation. So far, all previous research on VGI that used the IBM for positional accuracy applied the simplified version, while the iterative approach is not recorded to have been used until this research.

The Goodchild and Hunter's (1997) proposed iterative algorithm was tested and was proven unsuitable for the complex networks in this context, as compared to the single coastline that they used in their case study. Instead, a binary search algorithm is implemented in this study, which proved to be more efficient and robust and enabled the second, iterative and more complicated approach described in IBM.

4.12.2. The binary search algorithm

The following algorithm is applied on the matched reference and VGI datasets and works as follows:

1. The user is asked to define the desired overlap percentage (pct_{target}). A maximum number of iterations is also set to avoid infinite loops (e.g. i_{max}=20). Assuming an initial buffer b_0=0 and a corresponding overlap percentage p_0=0, a buffer width b_1 is selected for the first iteration (i=1). This is selected to be 8m, however the user can define a different value of buffer b_1 if necessary (see Appendix A).
2. The reference dataset is buffered and the length of the tested dataset that falls within the buffer is calculated and expressed as pct_{VGI}. It is then compared with the desired percentage pct_{target}. The process stops if: $| pct_{VGI} - pct_{target} |<0.25\%$, or if: $| b_i - b_{i-1}|<0.01$ m. These tolerances are selected after empirical search and considering the accuracy of the datasets involved, so that the number of iterations is reduced and more realistic results are produced. Smaller tolerances lead to successive buffers that differ less than 1 mm, which has no practical meaning for VGI. If the above conditions are not met, move to step 3.
3. The last three buffer values and last two percentage values are stored. The achieved overlapping percentage (pct_{VGI}) is compared against the target percentage (pct_{target}). Figure 4.10 describes the algorithm in pseudo-code and shows how the next buffer is calculated according to the percentage comparison. The reason of keeping three last buffer values is to cover cases of successive percentages of the same value (not shown in Figure 4.10 to avoid figure complexity). If this happens, the last buffer value is neglected and a new is calculated as the average of the previous two ones. Using the new buffer value, move back to step 2. If the maximum number of iterations is reached, move to step 4.
4. The buffer width that achieved the closest overlap percentage to the desired one is selected. In case of two such buffers (as a result of using rounded percentage values, e.g. when target percentage is 95% and buffer 10 m achieves 94.90%, while buffer 12 m achieves 95.10%), the smaller buffer is selected.

Using the thresholds of paragraph 2, the iterations needed for convergence barely reach nine (the examples of Table 4.4 demand between 6 and 8 iterations). Setting maximum number of loops to 20 is appropriate and can cover even lower thresholds of paragraph 2 (e.g. convergence to the target percentage better than 0.25%, or buffer accuracy better than 1 cm), however this will unlikely provide better results for VGI. It further justifies less reasonable choices of initial buffer. If, for example, the positional accuracy is 16.15 m for a tile but the user sets the initial

buffer to 0.1 m, 17 iterations will be needed to reach it, which also means additional computational time.

The suggested initial buffer of 8 m could also be 5, 6 or 10 m, which are considered reasonable values for VGI, and the algorithm works similarly. 8m is preferred, however, because its successive divisions and multiplications by two provide buffer values with no or few decimal digits (e.g. 8, 4, 2, 1, 0.5, 0.25).

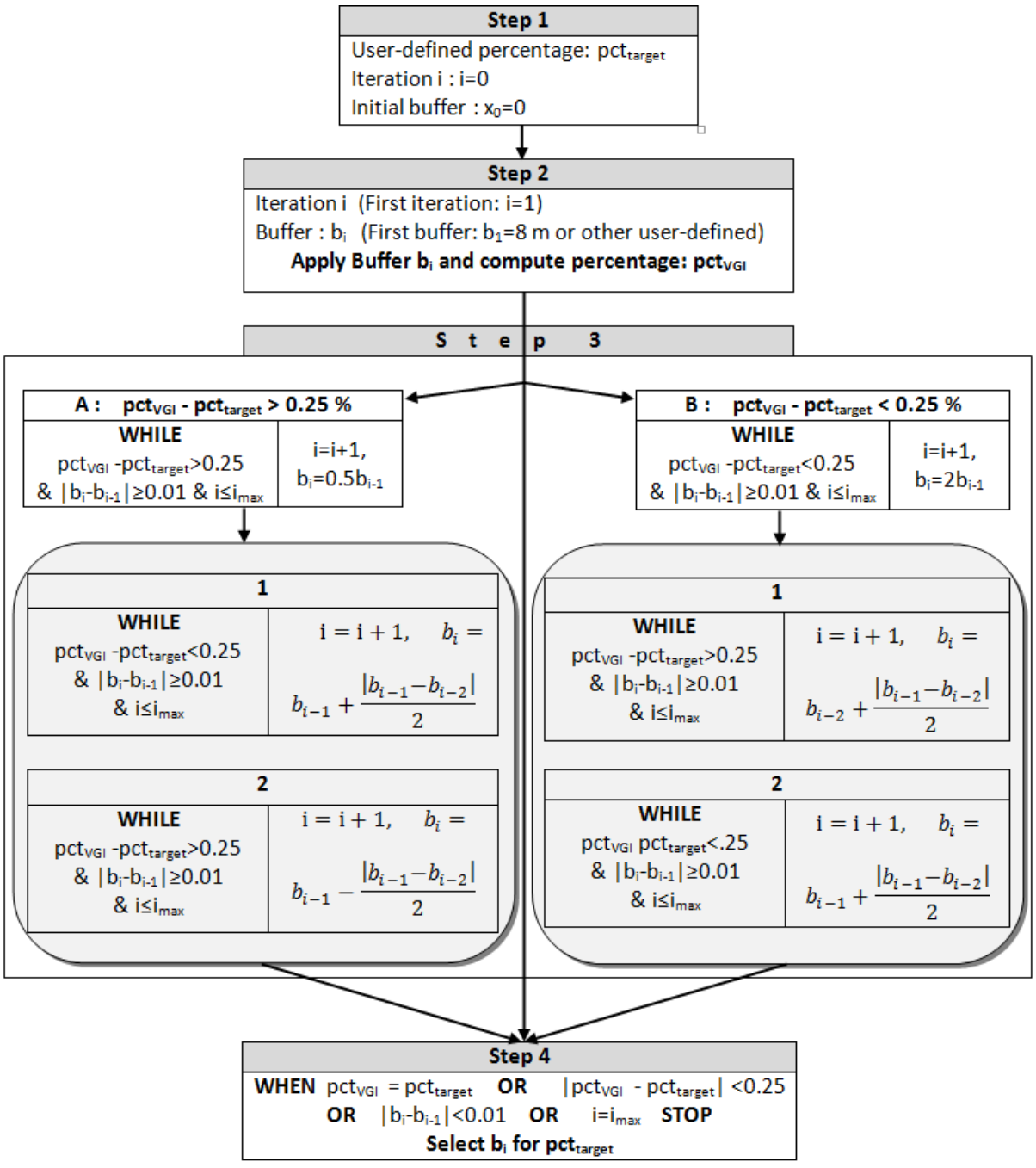

Figure 4.10: *The binary search algorithm*

Some examples of how the algorithm calculates the buffers are shown in Table 4.4.

Itera-tions	TQ4085		TQ4190		TQ5594		TQ5378	
	Buf. (m)	Pct. (%)	Buf. (m)	Pct. (%)	Buf. (m)	Pct. (%)	Buf. (m)	Pct. (%)
1	8	94.23	8	97.47	8	100	8	28.95
2	16	99.83	4	50.19	4	100	16	90.79
3	12	99.32	6	80.12	2	81.23	32	95.34
4	10	99.17	7	91.46	3	100	24	93.09
5	9	97.87	7.5	94.69	2.5	90.96	28	94.21
6	8.5	96.46	7.75	96.13	2.75	100	***30***	***94.77***
7	8.25	95.59	7.625	95.42	2.625	97.03		
8	***8.125***	***95***	***7.5625***	***95.05***	***2.5625***	***95.15***		

Table 4.4: *Examples of the binary search algorithm (target percentage 95%) for 4 tiles*

4.12.3. Positional accuracy assessment

The method is applied on the matched sub-datasets for each tile. The buffer is applied on the reference dataset. The results refer to each tile individually, succeeding in providing a localised positional accuracy value.

The user is free to decide on the desired overlap percentage. Section 8.4 argues that 95% is the highest 'safe' percentage and an adequate level of confidence. Percentages close to 100% should be avoided, because they lead to abnormally high buffer values for an increased number of tiles. These tiles are considered as 'outliers' and will be discussed in the next section.

Another reason for abnormally high buffer values could stem from objects close to the border. In Figure 4.11, corresponding reference object lies in the adjacent tile, which leads to an extremely high buffer value in order to include the VGI object that is left alone (as a whole or partially). This issue, however, is effectively handled using the extended tile (section 4.6); the buffer is applied on the extended reference sub-dataset, but the VGI sub-dataset examined is the one referring to the normal tile size. More details and examples are given in section 8.2.2.

Positional accuracy results for each tile are represented similarly to data matching results (explained in section 4.10 – Figure 4.9). Seven classes can also be used for positional accuracy values, however the buffer ranges of each class are decided accordingly in each case study.

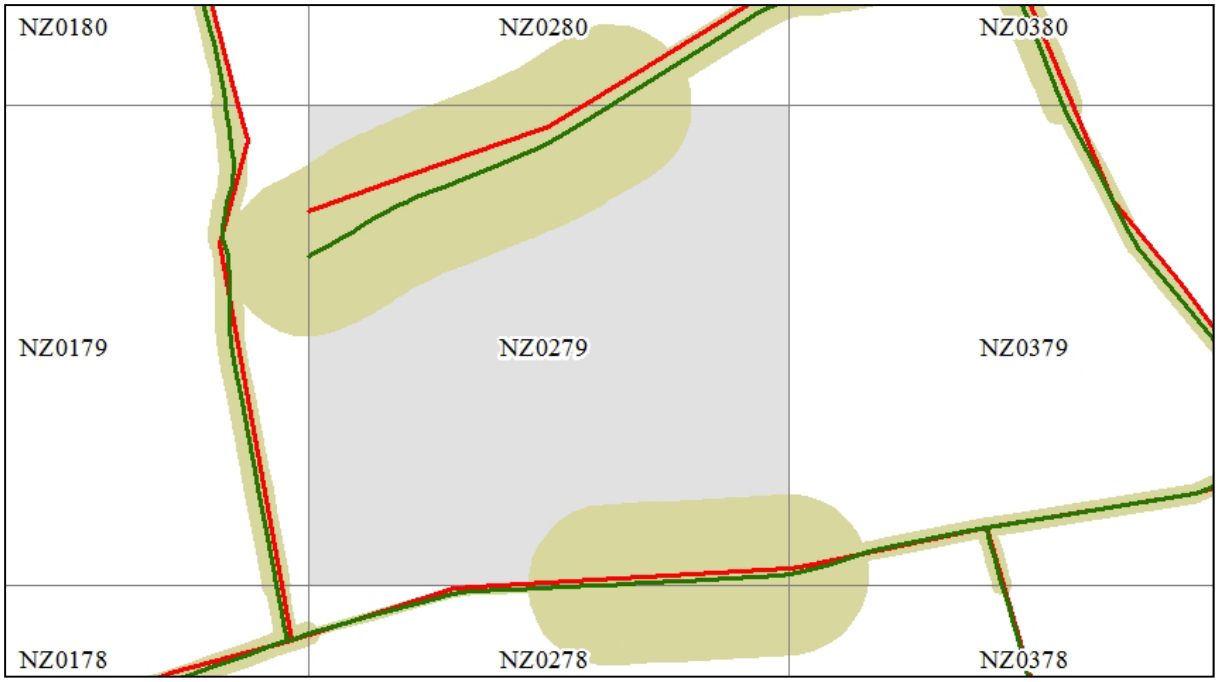

Figure 4.11: *Case of matched objects that demand a higher buffer value if extended tiles are not used*

4.12.4. Defining outliers

Section 4.8 mentioned that the data matching algorithm uses different maximum search distances for corresponding objects between urban and rural areas. Theoretically, any buffer width bigger than the maximum search distance could be considered as an outlier. In practice, however, this is not realistic for the following reasons:

- Data capture is different between datasets and corresponding objects are highly unlikely to be parallel in their full length. As shown in Figure 4.11, two corresponding objects may be within the search distance only partially. This is not a matching error; a buffer width bigger than the search distance would indeed represent the dataset's lower accuracy.
- The positional accuracy approach does not examine each feature individually, but refers to all existing data inside a tile.
- Data matching errors lead to the presence of features with no correspondence. Higher values of desired overlap percentage lead to a bigger buffer so that even these features will be included (even partially) inside the buffer. This also raises the question of a suitable desired overlap percentage; the choice should take into consideration the efficiency of data matching (Section 8.4 provides further analysis).
- Classification of urban or rural tiles depends on road density and the thresholds described in section 4.7 are decided based on tests in a specific rural area. In order to avoid data matching errors, the scale favours the urban areas. Slightly denser networks in rural areas result in classifying the tile as urban, leading to stricter matching constraints, however the

satellite imagery provided for VGI may have been of a lower resolution. Such cases demand a higher buffer value than the urban search distance which was used in data matching, so that to represent the lower positional accuracy of the tile more accurately.

After having performed some empirical tests, the threshold of classifying a buffer width as an outlier is set to roughly 1.5 times the maximum search distance; 40 and 75 m for urban and rural areas correspondingly. However, positional accuracy is calculated for all tiles and those considered as outliers simply obtain an additional 'outlier' attribute, so the above thresholds can be easily altered by the users, depending on their requirements.

4.13. Ending the process

4.13.1. Exported information

Section 4.9 mentioned that the resulting datasets from the data matching process are two for each source, one with the matched and one with the non-matched data. They are gradually created, populated with new data each time a tile is processed. These four spatial tables are exported at the end of the procedure. Other spatial tables, populated in the same way, are also exported as shapefiles and will be described in this section.

Additional output spatial datasets relevant to data matching but of minor importance are tables containing the matched and non-matched segments for each dataset. These were used during the development to test the matching procedure stage by stage and, as a result, they refer only to the first four stages, providing useful information when it comes to questioning a feature matching status.

The intersected VGI dataset (the one included in the final buffer, calculated using the desired overlapped percentage) is also exported, as well as the buffer polygon for each tile (Figure 4.11). This concludes the exported datasets that do not hold quality information referring to the whole tile. The processed tiles are exported as a tessellation file, which may differ from the one used as input if the datasets commonly refer to a smaller area.

The quality information that refers to each tile (data matching percentages, attribute accuracy, positional accuracy) is also stored gradually, this time on two non-spatial tables. One is for positional accuracy and the other one for all the remaining quality information. For the latter there are two

separate tables, one referring to the lengths and the other referring to the percentages (derived from the lengths). These are exported as Comma Separated Values (CSV) files in the end. Tables 4.5 and 4.6 provide two examples.

cell_id	dataset	Cell _pct	seg_ match	seg_no _match	seg_ match _1	seg_ match _2	seg_ match _3	seg_ match _4	seg_ ma_5 _add
TQ0580	OSM_urban	8802.38	5111.322	3691.058	1411.232	3077.013	191.558	431.52	31.106
TQ0580	ITN_urban	6219.14	5541.877	677.263	1175.56	3643.82	209.149	513.349	9.573
TQ5274	ITN_urban	2678.834	2498.19	180.644	830.833	1042.573	58.257	566.528	6.796
TQ5274	OSM_urban	3231.509	2495.503	736.006	1025.219	1132.368	53.05	284.867	0
TQ4992	OSM_urban	6290.282	4969.424	1320.859	1516.087	2017.812	0	1435.525	0
TQ4992	ITN_urban	5066.849	4722.045	344.804	1190.649	2400.88	0	1130.515	0

seg_ ma_5 _min	seg_ match _6	seg_ match _7	cell_ match	cell_no _match	rn_ comp _ini	rn_ comp _m	rn_ comp _nm	rn_ acc	ref_ acc	att_ acc
-34.383	13.903	18.431	5200.217	3602.163	25	24	1	4512.37	438.361	4950.731
-23.686	130.738		5622.235	596.905	25	25	0	5123.487	526.459	5649.946
0	87.325		2623.643	55.191	8	8	0	1492.999	193.973	1686.972
-64.079	27.054	0	2613.575	617.934	10	9	1	1638.82	191.078	1829.898
-64.814	0	0	5216.79	1073.492	24	23	1	3154.516	0	3154.516
0	0		4722.045	344.804	23	23	0	3657.539	0	3657.539

Table 4.5: *Example of the exported quality information (length table - no positional accuracy)*

gridcell	data	total_iterations	iteration	buffer	percentage
TQ0580		10	10	16.125	95.01
TQ5274		7	7	25	94.98
TQ0484		7	7	2.875	95.06
TQ5394		11	10	7.546875	95.38
TQ4992	outlier	9	9	108	94.92
TQ0491	outlier	10	10	49.5	95.12

Table 4.6: *Example of the exported positional accuracy information*

Separate tessellation files are produced, linking the non-spatial information of the CSV files with the spatial table of the processed tiles. Specifically, for each dataset there is one for data matching percentages, three for attribute accuracy percentages (for primary, secondary name and total attributes) and one for positional accuracy. There is also one for both datasets that provides the

mixed matching percentage (Figure 4.9d). These files provide the necessary tile information to deal with the VGI heterogeneity. Appendix A provides a detailed list of the exported tables.

Other descriptive information is also exported in CSV file format. Lengths used in quality information for each tile are summed up and a new table, following the same structure as Table 4.5, gives information for the whole dataset. Another table describes the buffer and overlap percentage achieved from each iteration of the positional accuracy evaluation and for each tile (Table 4.4). Road type correspondence between the datasets is also exported and discussed in the next section (Table 4.3 provides a partial example). Based on the road type and the lengths matched, information is also collected on what the two sources fail to map (in respect to each other). This is further explained for each case study in sections 5.5.3 and 7.5.3.

4.13.2. VGI commission: indication from road type correspondence

Information that extends tile borders and refers to the whole dataset is also processed at this stage. Specifically, this information concerns VGI commission. VGI commission is conceived as data not present in the reference dataset, but of a type compatible with the reference dataset specifications, so it should have been present. In other words, data updated in the VGI dataset but not yet added in the reference dataset.

Section 4.10 explained how road type correspondence information is aggregated. Although collected for each tile, it is more useful in the end when all the relevant information is gathered. For each reference road type, the prevailing VGI one is selected (for the example of 'B Road' in Table 4.3 it is 'secondary'). Moving to the VGI non-matched table, all features with this road type are marked. This means that these non-matched VGI features of the specific road type could be something that should exist in the reference dataset. This limits the search for VGI excessive data to feature types also collected by the reference dataset.

Due to the different classification of information between datasets (e.g. number of classes describing the road type) and lack of standards in VGI, it is not always possible to link classes between datasets in an absolute way. Table 4.3 provides such an example: a secondary road found with no match in VGI could correspond to many road types in the reference dataset. As a result, only indication on VGI commission can be given, which needs to be tested manually. Possible outcomes are that the VGI feature examined:

- is indeed additional information that should have been present in reference dataset, which is evidence of VGI commission.
- is not matched to its corresponding one in the reference dataset due to erroneous handling of the automated matching method, so it indicates data matching error instead of VGI commission.
- has wrong attribute value (e.g. footpath being classified as secondary road), therefore it is correctly assigned as a non-matched feature, but it is not VGI commission.

4.13.3. VGI commission: indication from attribute accuracy

Section 4.11.3 mentioned that non-matched VGI features with a road name attribute are marked to be checked for VGI commission. This is because an object with a name is less likely to be a footpath or other type of VGI not present in the reference dataset. Tagging usually demands field work, so the existence of a name means that the corresponding road exists and is of some importance to be labelled. This indication needs to be tested manually with the same outcomes as in the previous section.

4.13.4. VGI commission: indication from tile completeness results

The two previous sections described the indication on VGI commission at the feature level. A less efficient way of looking for excessive data in VGI datasets is by examining the completeness results. While 100% VGI matching percentage means that all features are found in the reference dataset, lower percentage values indicate VGI excess data. However this is a less efficient indicator because:

- it derives from all VGI road types, even those not designed for the reference dataset.
- in case of dense networks, a significant length of excess data within a tile is necessary to alter the matching percentage and visually alert the user.
- data matching errors influence the matching percentage.

The same applies to attribute accuracy percentages: lower VGI attribute accuracy percentages could be the reason of road names existing in VGI but not in the reference matched dataset.

4.14. Data matching evaluation, errors and impact on quality results

Data matching evaluation is performed manually for a data sample of the area tested, as will be described in each case study. Data matching errors influence quality elements measurement and may provide a higher (optimistic) or lower (pessimistic) value for the area examined.

Using the error levels found during the manual evaluation of data matching, estimation can be given on the errors regarding data completeness results. Data matching errors refer to:

- Reference objects mistakenly matched (surplus reference errors),
- Reference objects failed to be matched (missing reference errors),
- VGI objects mistakenly matched (surplus VGI errors) and
- VGI objects failed to be matched (missing VGI errors).

Although the total error for each data source is the sum of missed and surplus errors in terms of length, they compensate each other when it comes to calculating the total data completeness error. For VGI completeness, mistakenly matched reference objects lead to a bigger value of matched length, and successively to a higher matching percentage. Missed reference objects, on the other hand, reduce the matching percentage. As a result, the error in data completeness estimation can be generally calculated for the whole area according to equation 11 (using lengths):

VGI completeness error (omission) = surplus reference errors – missing reference errors (11)

If equation 11 returns a positive value, VGI completeness estimation is more optimistic and vice versa.

VGI attribute accuracy is less affected by data matching errors, however more unpredictably, despite also being calculated using the length of the reference features, as section 4.11 mentioned. The reason is that mistakenly matched or non-matched features do not necessarily have primary, secondary or both road names. A manual evaluation is also necessary for the matched data that will measure:

- Errors in data matching (error type 1), which lead to mistakenly matched objects with an attribute value but with no corresponding object to be compared.
- Errors due to a failed text similarity (error type 2), which accepts different names as similar ones because the similarity threshold (see section 4.11.2) is quite low for some cases.
- Errors due to a failed text similarity (error type 3), which rejects similar names because the similarity threshold is quite high for other cases.

For the total error estimation of attribute accuracy, error types compensate each other. For VGI attribute accuracy, error types 1 and 2 lead to a bigger value of matched length, and successively to a higher attribute accuracy percentage, while error type 3 reduces it. As a result, the error in attribute accuracy estimation can be generally calculated for the whole area according to equation 12 (using lengths):

$$\text{Attribute Accuracy error} = \text{ErrorType1} + \text{ErrorType2} - \text{ErrorType3} \quad (12)$$

Section 4.12 described how buffers are applied on all the matched reference features for the positional accuracy assessment. Reference features mistakenly considered with a match do not affect positional accuracy, because as there is no corresponding VGI object nearby, the buffer applied on them will not include additional VGI length to alter the positional accuracy result. Likewise, if VGI features are not considered as matched ones, they will not be included in the matched subset and will not be examined. The opposite, however, affects positional accuracy when the corresponding VGI object is considered as matched but not the reference one. Depending on the desired overlap percentage, if this VGI feature has to be included inside the buffer but the corresponding reference one is not used (by not being found as matched), the buffer on the nearest reference feature (but not the corresponding one) will have to grow significantly in order to include it, which gives a much lower positional accuracy value than the actual one for the tile under examination. The advantage of the positional accuracy approach in this study is that the results will never be higher than the actual ones, leading to safer decisions, however much lower results give the wrong impression of VGI quality. Equation 13 calculates the amount of errors in data matching that will likely affect positional accuracy.

$$\text{Data affecting pos. accuracy} = \frac{\text{missing Ref.errors (m)} + \text{surplus VGI errors (m)}}{\text{Ref.length examined (m)} + \text{VGI length examined (m)}} \% \quad (13)$$

However, since positional accuracy results do not refer to the feature level but to the whole tile, an average positional accuracy error cannot be calculated. The effect of erroneous data matching on positional accuracy can be indirectly calculated by examining a sample of outliers. The number of outliers caused by data matching errors, compared to the total number of sampled tiles, gives an estimation of the total tiles that have their positional accuracy affected by data matching errors (equation 14).

$$\text{Positional accuracy errors} = \frac{\text{Number of outliers caused by errors in data matching}}{\text{Total number of tiles}} \% \quad (14)$$

More details about the manual evaluation are supplied in sections 5.4, 6.4, 7.4 respectively for each of the case studies used in this thesis.

4.15. Justification of parameters and options

Tables 4.7 and 4.8 present the parameters that were used in the proposed framework, informing where further justification can be found for those that have not been fully explained during the method description. Some of them are the result of tests during the case studies. Since they described using results, table structures and presentation techniques that the reader is not yet familiar with, they are presented in Chapter 8 (discussion chapter).

	Parameter	Value	Short description	Appeared in Section	Explained in section
1	Tile shape and size	Square, 1 km^2	Areal unit to represent VGI heterogeneity	4.6	4.15.1, 8.2.1
2	Buffer width to extend tile	50 m	Deals with data matching close to the tile border	4.6	4.8.3, 8.2.2
3	Tile classification as rural	17 feat./km, 8 jun./km	Threshold to classify a tile as urban or rural	4.7	8.2.3
4	GPS accuracy a	Urban: 10m Rural: 15m	Parameter that defines search distance and angular tolerance	4.8.3	4.15.2
5	Rural integer c	Urban: 2 Rural: 3	Parameters that define search distance for corresponding objects between datasets	4.8.3	4.15.2
6	Road width w	2 – 11 m		4.8.3	4.15.2
7	Angular tolerance	φ	Defines which segments are considered to be parallel	4.8.3	4.8.3
8	namematch1	65%	Threshold for text similarity of stage 3 (data matching)	4.8.5	4.8.5, 8.5
9	namematch2	75%	Threshold for text similarity of stage 6 (data matching)	4.8.8	4.8.8, 8.5
10	Stage 7 data matching parameters	0.8 0.9	Deal with cases of generalised VGI data (shortest features) in data matching	4.8.9	4.8.9

Table 4.7: *Applied values for the parameters used during data matching approach*

	Parameter	Value	Short description	Appeared in Section	Explained in section
10	namematch3	70%	Threshold for text similarity of stages 2 and 3 (attribute acc.)	4.11.2	4.11.2, 8.5
11	Search distance for attribute acc.	Urban: 35m Rural: 50m	Area within which features with similar names are sought	4.11.2	4.11.2
12	Maximum loops	20	Max. number of iterations for positional accuracy	4.12.2	4.12.2
13	Initial buffer	8 m	First buffer for positional accuracy	4.12.2	4.12.2
14	IBM tolerances	Pct: 0.25% Dbuf: 0.01m	Define when IBM is considered to have converged to a result	4.12.2	4.12.2
15	Target percentage	95%	Level of confidence for positional accuracy results	4.12.3	8.4
16	Positional accuracy outliers	Urban: 40m Rural: 75m	Maximum buffer value accepted as positional acc.	4.12.4	4.12.4

Table 4.8: *Applied values for the parameters used during attribute and positional accuracy*

4.15.1. Tile size and shape

As mentioned in the literature review chapters, one way of dealing with VGI heterogeneity is to produce results for smaller areas. This can be achieved using a tessellation file, which is a collection of adjacent polygons, where each polygon can be used to clip the datasets. As a result, each tile is processed individually and local measurements of data quality are possible. Additionally, computation is faster by limiting the number of objects processed each time. There are, however, two things to consider:

Tile shape: Polygons of different shapes can be used, such as administrative boundaries. Although results can be more meaningful when referring to areas well-defined on the ground, the size may still be too big to deal with VGI heterogeneity, especially in mixed rural and urban areas. Additionally, since administrative boundaries are usually defined by roads, it may occur that the same object described in two different datasets will lie in different tiles, despite the objects' close distance (Figure 4.12). In such cases a tile-by-tile examination may include errors close to the tile borders. A normalised grid, where each tile has the same size, seems to be more efficient for both

the problems described; size remains the same and results refer to equal-sized areas, while the number of objects close and parallel to the tile borders is reduced. Section 4.6 described how this can be further eliminated with the use of extended tiles. As a result, a normalised grid will be used in all case studies, so that results will refer to equal areas.

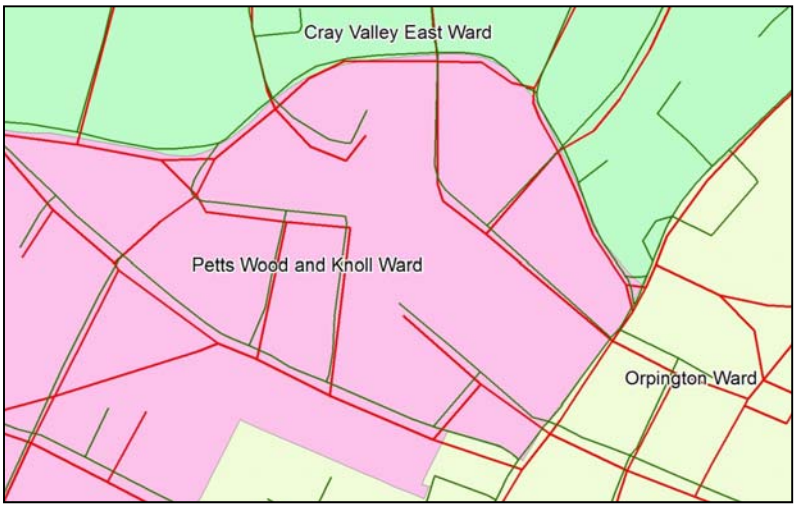

Figure 4.12: *Problems with corresponding roads when using administrative boundaries as tiles*

Tile size: The next concern is the tile size of the grid. The size has to be relatively small for computational reasons (less data to process each time), but at the same time practically meaningful. Since VGI can refer to data collected by an individual on the field, a tile size of 1 km^2 seems a reasonable area to be mapped by one person, even in a dense urban network. Haklay *et al.* (2010) also use the same spatial unit, suggesting that larger tiles would be less appropriate to describe spatial quality due to the VGI heterogeneity.

Grids of 1 km^2, 5 km^2, 10 km^2, 100 km^2 are available from the OS for the UK, however they can only be applied in the UK. The OS's 1 km^2 National Grid was used for the first two case studies, so that results would refer to an official grid and facilitate a potential cross-examination. Using Manifold and ArcGIS software, a 2 km^2 tessellation file was additionally created for the first case study (used in section 8.2.1) and a 1 km^2 file for the third case study (Haiti area).

4.15.2. Search distance parameters a,c,w

Section 4.8.3 described the parameters a,c,w that define the search distance, within which corresponding features are sought during the matching procedure. Unfortunately, the differences between data sources do not permit an optimisation of their values that will render them appropriate for all cases. The selected values (Table 4.7) generally proved sufficient for the case studies used in this thesis. Section 4.8.3, although justifying the tile classification thresholds, shows

that data matching is affected while altering these values. The urban area tested (Greater London) includes data of high density, completeness and positional accuracy from both reference and VGI datasets. The efficiency of the matching algorithm with such networks, as will be proved, implies that these are the lowest values to be used, as datasets from other sources are unlikely to have even better quality (which might require stricter constraints – lower values). In case of datasets of lower quality, these parameters will need to be modified (increased) by the user, which would provide a greater searching distance within which corresponding objects are sought. The automation of the method enables the user to run it with sample data, so that these parameters can be manually adjusted. Otherwise, the use of stricter constraints when looser ones are needed, will increase the percentages of erroneous data matching.

Having 3 parameters to define search distance, allows for a better customisation. Road widths (w) may differ between countries, areas or datasets, so the user should be able to alter them. The 'rural integer' (c) allows for a bigger (or smaller) difference between urban and rural searching distances, if necessary. GPS accuracy (a), finally, depends on the device.

4.16. Areas and datasets for the three case studies

The framework that was described here was applied in 3 case studies to evaluate its efficiency and effectiveness. This section briefly presents the selected areas for each case study.

The first case study involves two areas in the UK, specifically the Greater London and a rural area west of Newcastle. Chapter 5 provides more information on the selection of these areas, along with the results and evaluation of the method.

The second case study extends the areas of the first to cover England and Wales region by region, in order to check the method in far bigger areas with a mixed network type in terms of density and accuracy (urban and rural). Chapter 6 provides more information on the analysis and results.

The third and final case study checks how generalized the proposed framework is, by examining two totally different sources and areas. Specifically, the studied area is the capital city of Haiti, where official data are provided by NATO (MINUSTAH project). From the VGI point of view, Google Map Maker is used as well as OSM. Additionally, a comparison between the two VGI datasets is performed to check if the framework horizons can be broadened to cases where no official reference dataset is available. These are further elaborated in Chapter 7.

4.17. Summary

The proposed methodology tackles VGI quality in an automated and systematic way that accepts heterogeneity, lack of standards and lack of uniform density in objects or their attributes. Quality elements of data completeness, attribute and positional accuracy are measured by comparing linear VGI with a reference source of known quality, such as official datasets from a mapping agency. The focus is first on finding the unique objects on each dataset, which are then removed, leaving datasets with data describing the same objects.

The removed objects form a different dataset that can be used for conflation purposes or for finding commissioned data (in other words excess data that should be present in one dataset but are not, while the other seems updated). The focus then moves to the remaining corresponding objects, which also form different datasets that are further processed to measure the above mentioned quality elements.

Results are produced for each tile of a tessellation file. The evaluation of all quality elements is based on measuring the length of the road network, which seems to be a more useful indicator for linear data and leads to results more representative of the phenomenon that they aim to describe. Data completeness uses the length of matched data as compared with the total length for each tile. While results refer to the tile level, information is gathered during the matching procedure and stored in each feature as additional attribute, providing a more detailed picture of data completeness, especially for commissioned data. Attribute accuracy and completeness also rely on features length, looking for similar primary and secondary names. Again results refer to the tile level, although relevant information is stored also at the feature level. Positional accuracy, finally, uses increasing buffers to calculate the length of VGI included in a specific buffer size. In this case, however, results refer only to the tile level.

Three different case studies were selected. They differ in data density, uniformity, accuracy and types of information they provide. Different data sources are also used to check how generic and robust the proposed method is. The efficiency level of the method in these cases aims to test if this framework can generally provide the necessary quality information for someone to decide whether a VGI source is suitable for specific requirements, enabling potential usage of VGI, as described in sections 1.5 and 1.8.

Chapter 5

First Case Study: Urban and rural area

5. First case study: Urban and Rural area[5]

5.1. Introduction

After describing the framework and briefly presenting the three case studies in Chapter 4, this chapter focuses on the first case study, which includes the Greater London and a rural area west of Newcastle. OSM and ITN datasets are compared as VGI and reference datasets correspondingly. This analysis includes justification of the study area, application of the method, results, evaluation and discussion.

Section 2.3.1 presented the OSM project, which will be used in this case study as a VGI source. The next section briefly describes the ITN dataset, which will be used as a reference data source.

5.1.1. Reference Data Sources: Ordnance Survey's MasterMap - ITN Layer

OS's Integrated Transport Network (ITN) is selected for the first two case studies as a reference dataset. ITN is one of the OS Master Map Layers, consisting of '*a fully topologically structured link-and-node network representing the roads network of Great Britain, from motorways to pedestrianised streets'* (Ordnance Survey, 2009a, p.16). Its accuracy is specified as the accuracy of OS Mastermap Topography Layer, derived from mapping sources of scales 1:1,250, 1:2,500 and 1:10,000 for Urban, Rural and Mountain / Moorland areas respectively, resulting in accuracies of 1.0 m, 2.5 m to 6.0 m and 8.0 m accordingly (Ordnance Survey, 2009c, p.80). While routing information is updated within six months of a potential change in the real world 'wherever possible', an update of the data is available every six weeks. Tracks or paths that are not driveable by an ordinary vehicle (e.g. a family car) are not included in the ITN layer. It is the most detailed road network provided by the OS and the most accurate official data covering the whole country. Road network is represented by linear features, drawn along the axis of a road. Ordnance Survey (2009b) describes all the attributes assigned to ITN features in details. Roads are classified in 9 groups, however it is mentioned (p.17) that the classification is not cross-referenced with third parties.

Data download is allowed in UK for academic and research purposes through an institutional licensing framework.

[5] Sections 5.2, 5.3, 5.4.1, 5.4.2, 5.5.1, 5.5.2, 5.5.3 have been partially adapted from:
Koukoletsos, T., Haklay, M. and Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, [in press - DOI: 10.1111/j.1467-9671.2012.01304.x].

5.2. Area justification and data preparation

Although both areas include sub-areas with dense and scarce networks, as a result of the existence of built-up areas, Greater London will be treated as an 'urban' and the area west of Newcastle as a 'rural'. Sections 4.7 and 8.2.3 include a more in-depth discussion of this classification.

The reason for selecting the Greater London is to test the method in an area which has already been partially studied and found to have good VGI quality in terms of completeness and accuracy, so that results could be compared. Additionally, the efficiency of the method of automated data matching will be tested in an area where both VGI and reference datasets are of high density. The rural area, on the other hand, is selected in order to test how the methodology applies to an area away from where OSM started, where networks are not as dense, VGI is considered less complete and positional accuracy is reduced for both reference (see section 4.16.4) and VGI datasets (Haklay, 2010c).

The OSM dataset covering the UK was downloaded on 21/10/2010 in shapefile format from geofabrik (2010). The corresponding ITN dataset was obtained through the institutional licensing framework at the same time, so that the temporal conditions between the datasets remain the same. Data for the urban area of Greater London (1,731 km^2) were clipped using the corresponding district boundary, also obtained from OS. The shapes of study areas are a result of following the district boundaries (Figure 5.1a). For the rural area west of Newcastle, though, district boundaries were used for the south and west borders, while north and east borders follow straight lines to avoid including built-up areas (with a denser network) as well as to reduce the size of the tested area to a size close to the urban one (2,120 km^2, Figure 5.1b). The tesselation used is the 1 km^2 National Grid. Table 5.1 provides information on the total network length. Generally, most of the following tables keep a similar structure for all case studies: each row refers to the studied area (e.g. Urban) and is further divided into two rows, one for each dataset examined (e.g. ITN and OSM).

Area Description	Type	Total Area size	Compared Area size	Data set	Total network length (m)	Average length (m) per tile
Greater London	Urban	1,731 km^2	1,687 km^2	ITN	18,368,381	10,611
				OSM	20,229,408	11,686
West of Newcastle	Rural	2,120 km^2	1,299 km^2	ITN	2,935,672	1,385
				OSM	1,872,836	883

Table 5.1: *Studied areas and road network information*

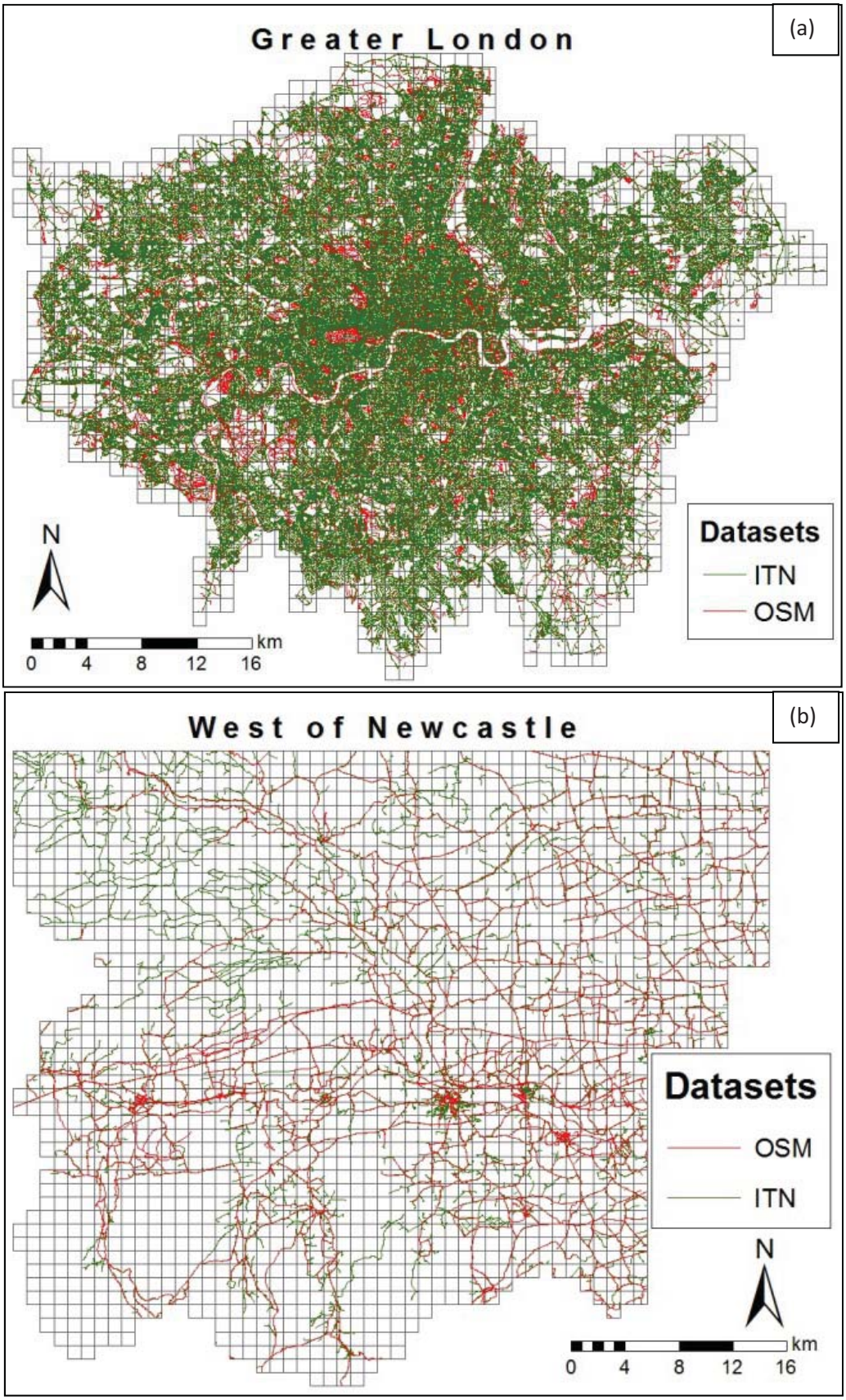

Figure 5.1: *Areas, datasets and tiles for the 1st case study:* ***a.****Urban area,* ***b.****Rural area*

'Compared' tiles refer to areas where both datasets contain data, so a different tiling method (tile size or position) may lead to a different number of actually compared tiles. Using the 1 km^2 tessellation file, for example, this results in 1,687 and 1,299 km^2 for the urban and rural areas accordingly, described as 'compared area size' in Table 5.1. Section 8.2.1 discusses further the use of different tile sizes. The average length is the total dataset length divided by the total number of tiles. The difference between the 'total' and 'compared' number of tiles or area size in rural area explains the low average length values of Table 5.1.

ITN and OSM datasets have different coordinate systems. The first uses the British National Grid (BNG, projected), while the second uses WGS84 datum (latitude-longitude projection). As ITN is the reference dataset, the analysis was carried out using the BNG, so OSM had to be reprojected. This is to avoid any distortion on the reference dataset, as a result from the reprojection tolerances. Datasets are then loaded into the PostGIS database using Quantum GIS open-source software.

5.3. Method application and results

Using the application that was developed to enable interaction with the datasets and perform the analysis (described in Appendix A), each area is processed individually, following the flow diagram of Figure 4.1. Figures 5.2 and 5.3 present the output datasets for the urban area. Non-matched datasets are presented in one figure (Figure 5.4), because they refer to data unique in one place, so generally they do not overlap when viewed.

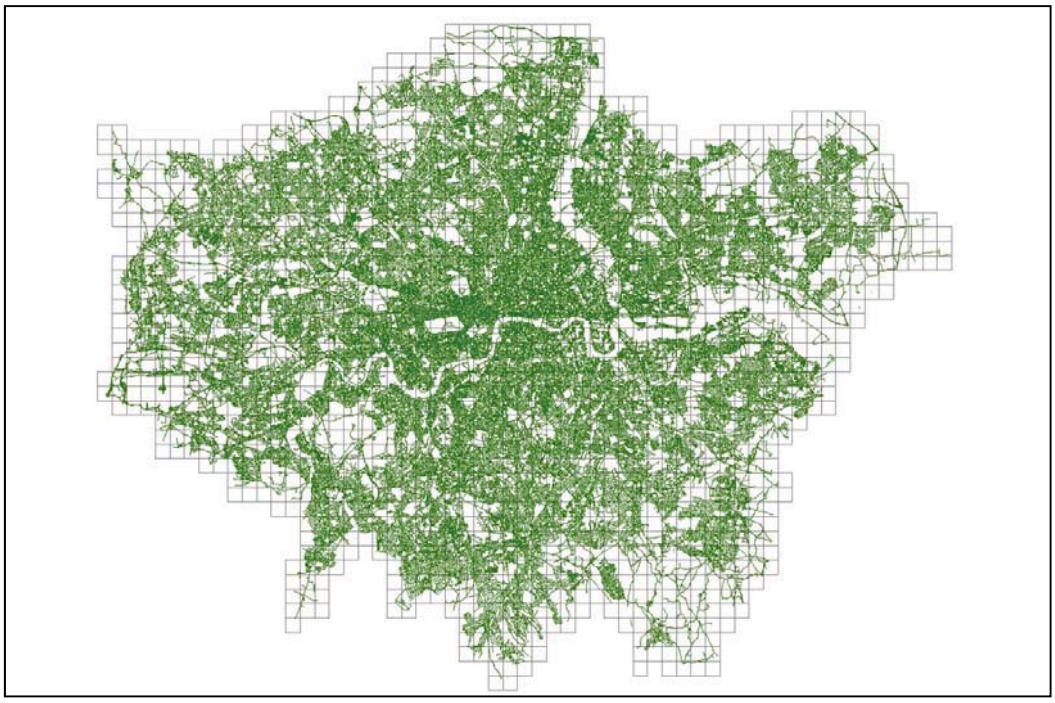

Figure 5.2: *Urban area: Matched reference (ITN) dataset*

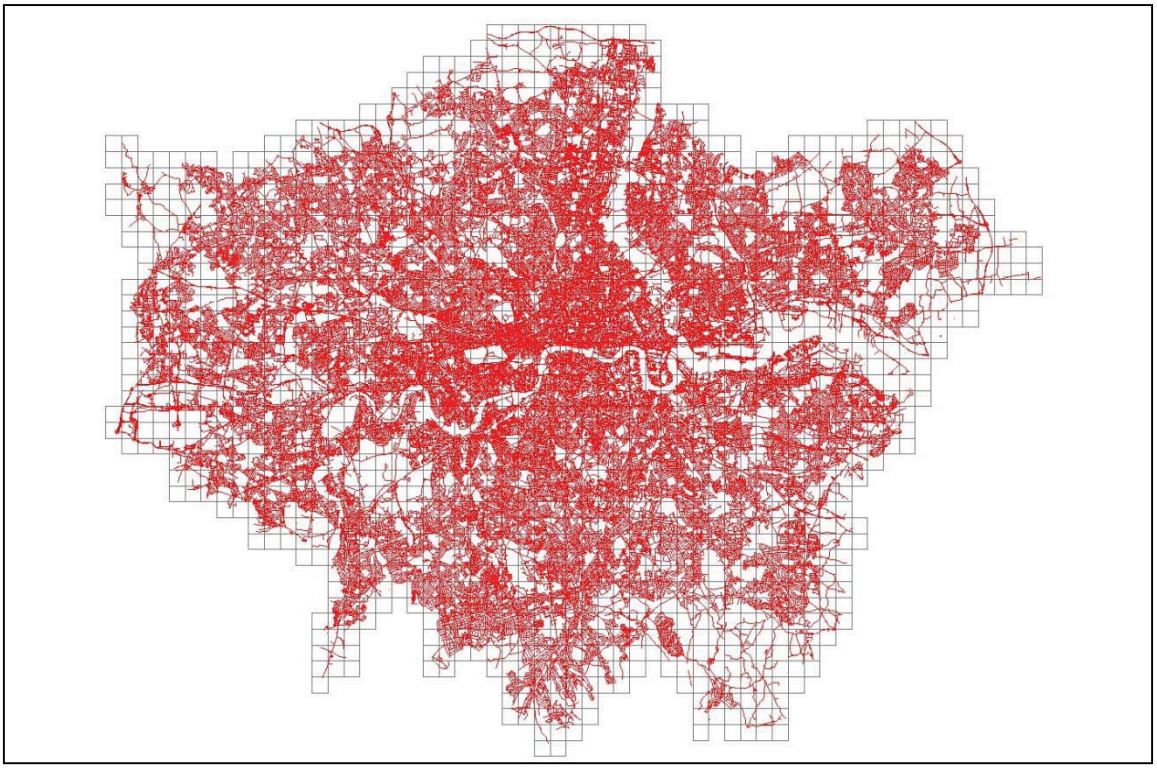

Figure 5.3: *Urban area: Matched VGI (OSM) dataset*

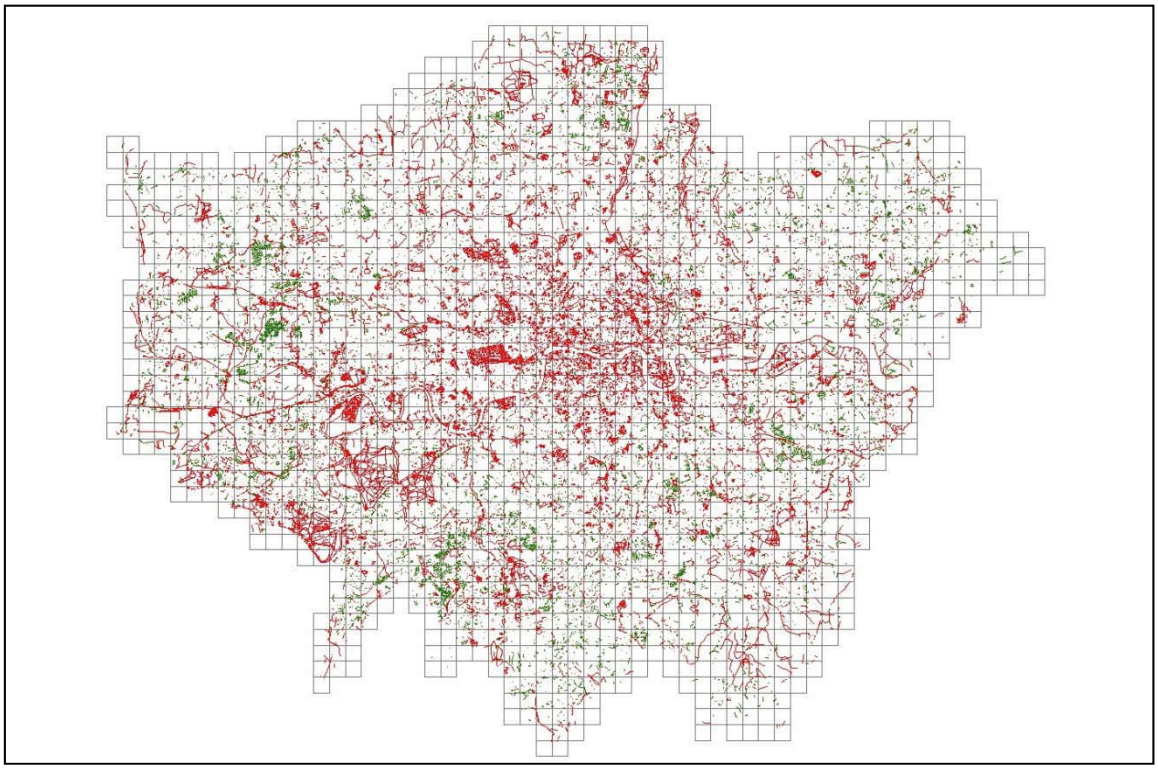

Figure 5.4: *Urban area: Non-matched VGI (OSM - red) and reference (ITN - green) datasets*

Similarly, Figures 5.5 and 5.6 present the output matched datasets for the rural area, while non-matched datasets are presented in Figure 5.7.

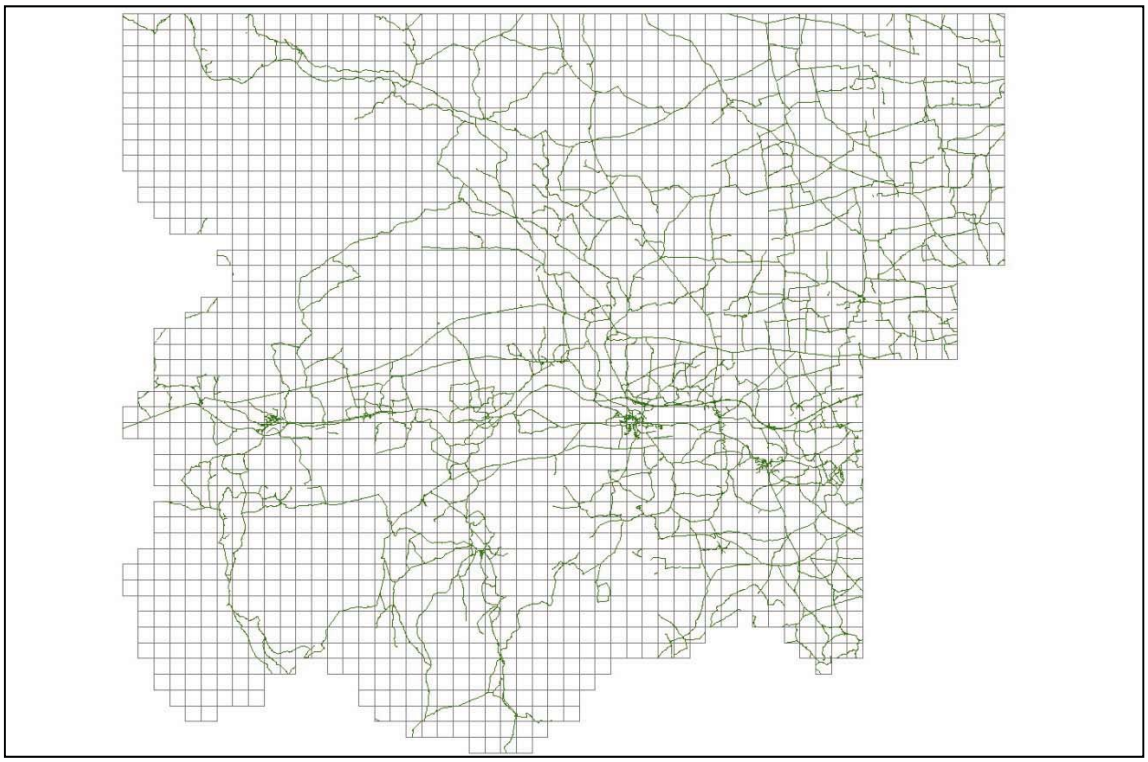

Figure 5.5: *Rural area: Matched reference (ITN) dataset*

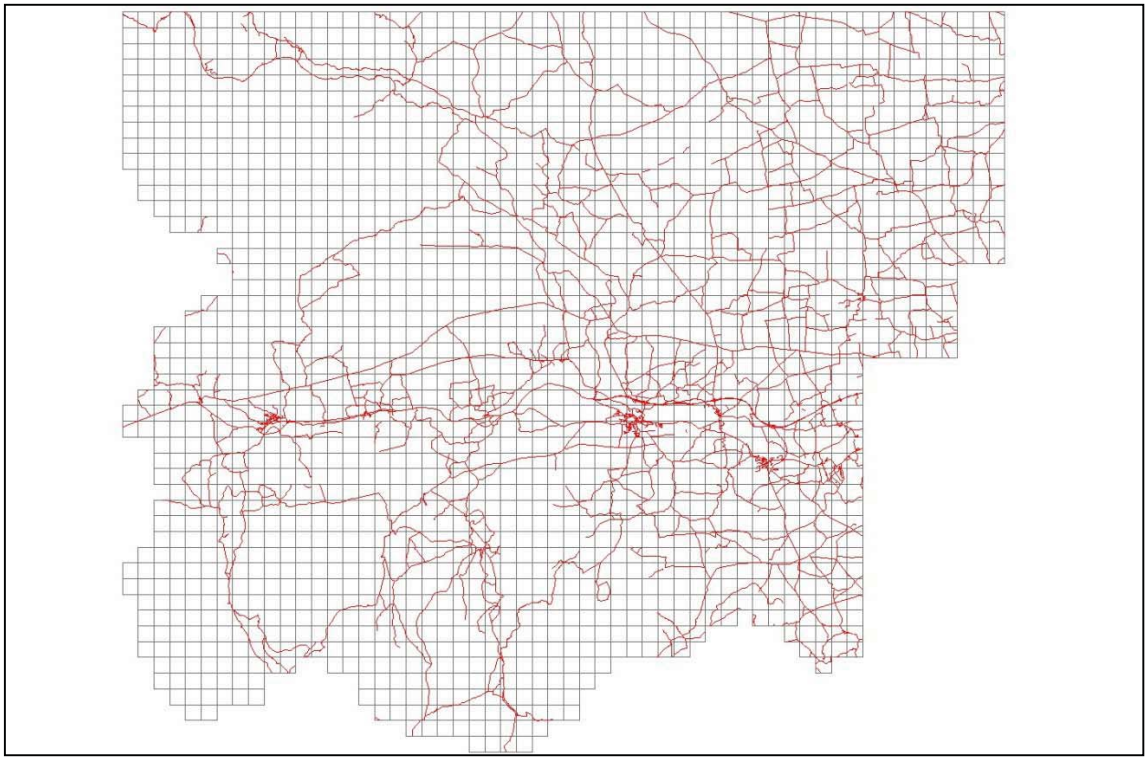

Figure 5.6: *Rural area: Matched VGI (OSM) dataset*

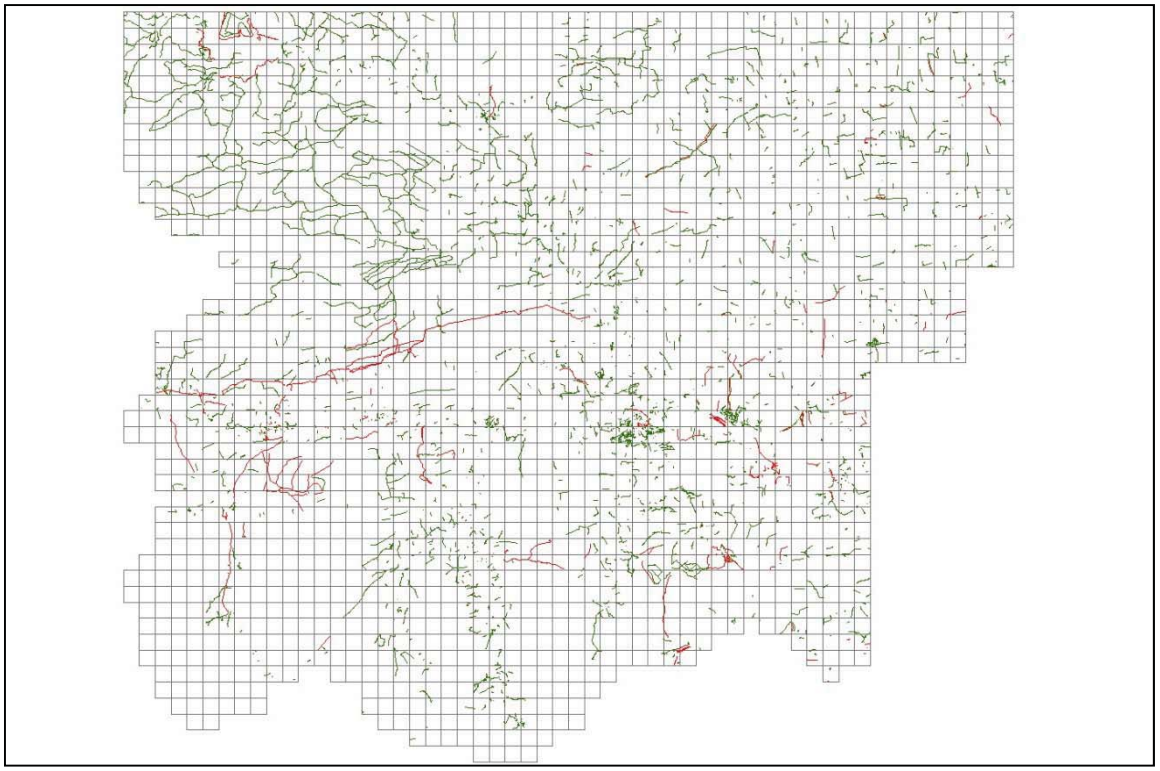

Figure 5.7: *Rural area: Non-matched VGI (red) and reference (green) datasets*

Table 5.2 presents the total lengths for both areas. Since the method is applied on the tiles where data from both datasets are present, 'Length Compared' is less than 'Total Length'. (Detailed CSV files that describe each tile individually are produced along with the relevant shapefiles).

Area	Dataset	Total Length (m)	Length (m) Compared	Length (m) Matched	Length (m) non-matched
Urban (Greater London)	ITN	18,368,148	18,366,935 (99.99 %)	17,084,539 (93.01 %)	1,283,609 (6.99%)
	OSM	20,734,150	20,719,274 (99.93 %)	16,718,721 (80.63 %)	4,015,429 (19.37%)
Rural (west of Newcastle)	ITN	2,935,675	2,500,826 (85.19 %)	1,735,695 (59.12 %)	1,199,980 (40.88%)
	OSM	1,952,632	1,922,656 (98.46 %)	1,719,435 (88.06 %)	233,197 (11.94%)

Table 5.2: *Resulting network lengths for study areas*

Figures 5.8 to 5.11 present the quality results for each tile for the urban area. Specifically, Figure 5.8 refers to data completeness (see section 4.10 for more details on the results calculation) and positional accuracy. A 'greener' view of ITN matching percentages (VGI completeness), compared to the OSM ones, means that more ITN objects are found in OSM dataset than the other way around, in

other words OSM contains much more additional data not present in the ITN dataset. OSM is more complete in the centre of London compared to the suburbs, demonstrated by ITN matching percentages between 90 and 99%. Additionally, OSM also contains additional features not present in the reference dataset in the centre of London, with OSM matching percentages generally less than 90%. This information usually refers to cycleways, steps, footpaths crossing the parks and other data types not present in the reference dataset. However, OSM and ITN's level of agreement (mixed matching perentages) is generally above 75%. OSM positional accuracy, finally, seems to be quite uniform (8-12 m for 72% of the tiles), while outliers reach 2.8% of the total tiles.

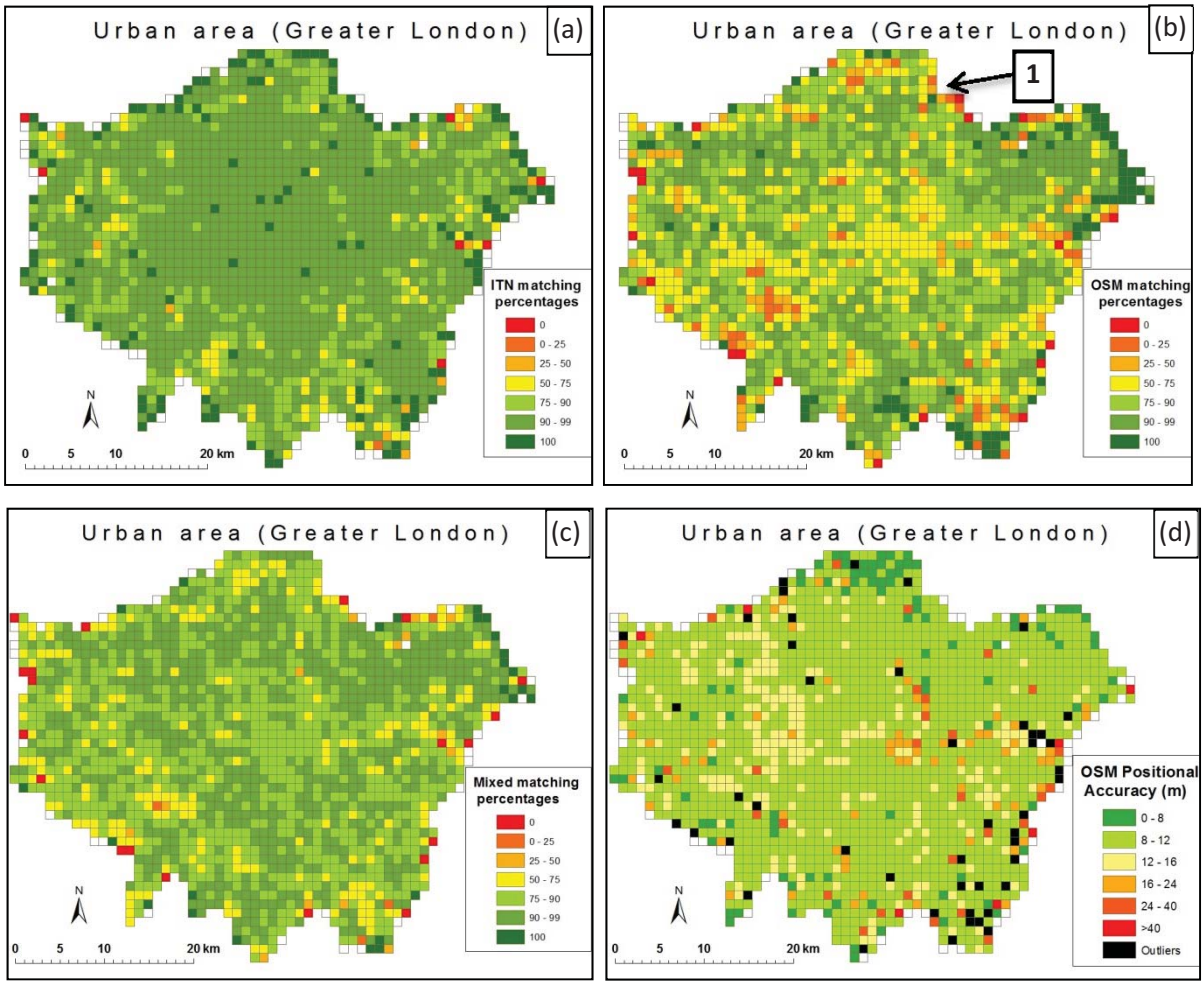

Figure 5.8: *Urban area – data completeness and positional accuracy,* ***a:*** *ITN matching percentages (OSM completeness)* ***b:*** *OSM matching percentages (OSM commission)* ***c:*** *Mixed percentages (level of agreement between datasets)* ***d:*** *Positional accuracy*

Figures 5.9 to 5.11 refer to attribute accuracy of primary, secondary and both names respectively in urban area. Results are presented in the same way as in previous figure. A greener view in OSM figures (right) compared to ITN (left) show that ITN is richer in attributes, in other words there are

road names in the ITN dataset not found in OSM. Specifically, primary names in OSM generally seem to be complete and accurate above 90% (Figure 5.9a), with the exception of the south-east area. OSM secondary names seem to be less complete in the centre of the area (Figure 5.10a), while lower OSM percentage values in the east (Figure 5.10b) indicate that there are additional attributes not present in the ITN dataset. This needs to be further examined to find out if it is a case of VGI commission or a systematic error in VGI, maybe attributed to a specific user. Total attribute accuracy, finally, proves to be quite high (Figure 5.11), mostly above 90%.

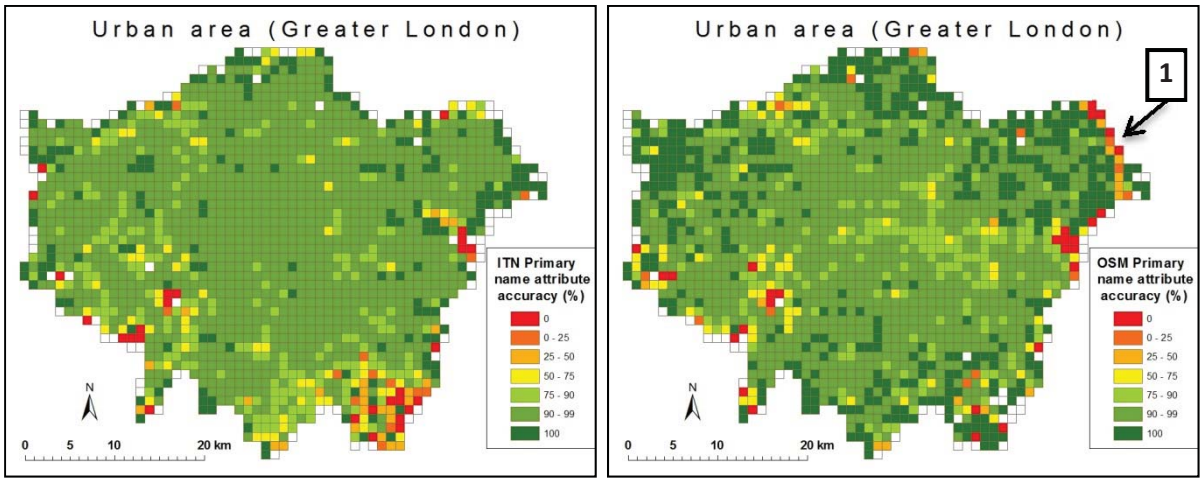

Figure 5.9: *Urban area – primary name,* ***a:*** *ITN percentages (OSM attribute accuracy)* ***b:*** *OSM percentages*

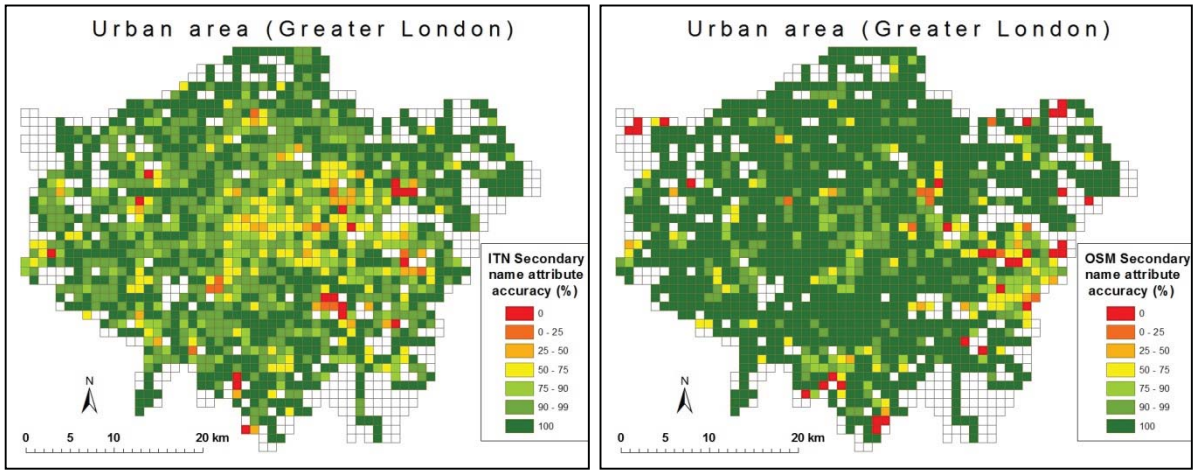

Figure 5.10: *Urban area – secondary name,* ***a:*** *ITN percentages (OSM attribute accuracy)* ***b:*** *OSM percentages*

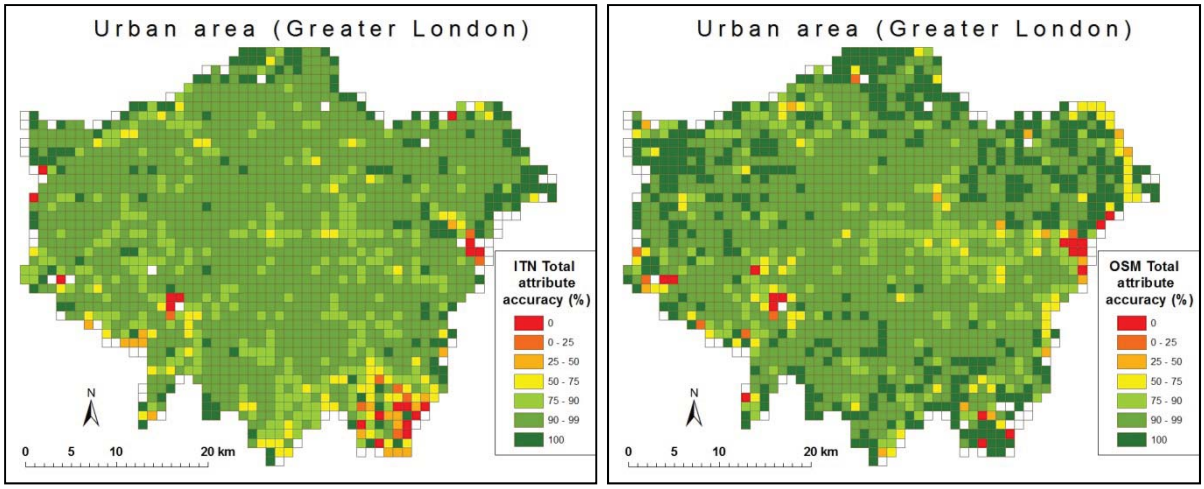

Figure 5.11: *Urban area – total names attribute accuracy,* ***a:*** *ITN percentages (OSM attribute accuracy)* ***b:*** *OSM percentages*

Figure 5.12: *Rural area – data completeness and positional accuracy,* ***a:*** *ITN matching percentages (OSM completeness)* ***b:*** *OSM matching percentages (OSM commission)* ***c:*** *Mixed percentages (level of agreement between datasets)* ***d:*** *Positional accuracy*

Moving to the rural area, Figures 5.12 to 5.15 present the quality results. Specifically, Figure 5.12 refers to data completeness and positional accuracy in a similar way to the urban area, however the greener view now is in OSM matching percentages compared to the ITN ones, which means that ITN is richer than OSM in terms of mapped objects. OSM generally does not contain additional features, compared to ITN, which is the reason why in Figure 5.12b OSM matching percentages are high (mostly 100%). OSM is also only sporadically complete, which is the reason why ITN matching percentages (Figure 5.12a) are generally lower or even zero. Their level of agreement (mixed matching percentages – Figure 5.12c) has a broader range of values than in the urban area (Figure 5.8c). OSM is also less homogeneous and accurate in terms of positional accuracy, while outliers reach 5.75% of the total tiles (Figure 5.12d). All figures demonstrate a less homogeneous OSM dataset than in the urban case.

Tiles with no values (blank or white ones) need to be explained further. They generally indicate that the examined dataset does not contain the information that is measured in each figure. This does not necessarily mean that the other dataset will also be without the relevant information, and that is why the same tile may be with quality value in one dataset but not in the other one. This is more likely to happen in rural areas, hence a more detailed explanation is given here.

If one of the datasets does not contain any features in one tile, this will be left blank (no value) when examining data completeness. An example is the north-west tiles of the rural area, where ITN dataset is richer (Figure 5.7). ITN matching percentages (Figure 5.12a) are zero because there are ITN features but no OSM features (or at least not found as corresponding), which effectively shows OSM completeness (which is zero) as explained in section 4.10. OSM matching percentages for the same tiles, where there are no OSM features, are left with no value, so as to differentiate from the case where features may exist but with no match in the other dataset, which would be the case of zero value. Cases where both datasets have data but nothing is matched are represented by zero matching percentages in both datasets (cases of the same red tiles in both Figures 5.12a, 5.12b). Mixed matching percentage (that shows the level of agreement between datasets) refers to tiles where non-null values exist at least in one dataset. For positional accuracy null values may refer to tiles where there are no VGI features or to tiles where no matched reference segments are found. In the second case, even if there are VGI features, there is no way to evaluate their positional accuracy, as there is nothing to compare them with.

Figures 5.13 to 5.15 refer to attribute accuracy of primary, secondary or both names respectively in rural area (similarly to 5.9-5.11 for the urban area). Here, however, primary names are scarce in the ITN dataset and generally non-existing in the OSM dataset (Figures 5.13a, 5.13b). Secondary names, however, are more complete and accurate (Figures 5.14a, 5.14b), reaching 100% for most of the tiles. The combination of Figures 5.13 and 5.14 leads to the total attribute accuracy of Figure 5.15, which shows that OSM is much more heterogeneous than (and not as complete as in) the urban area.

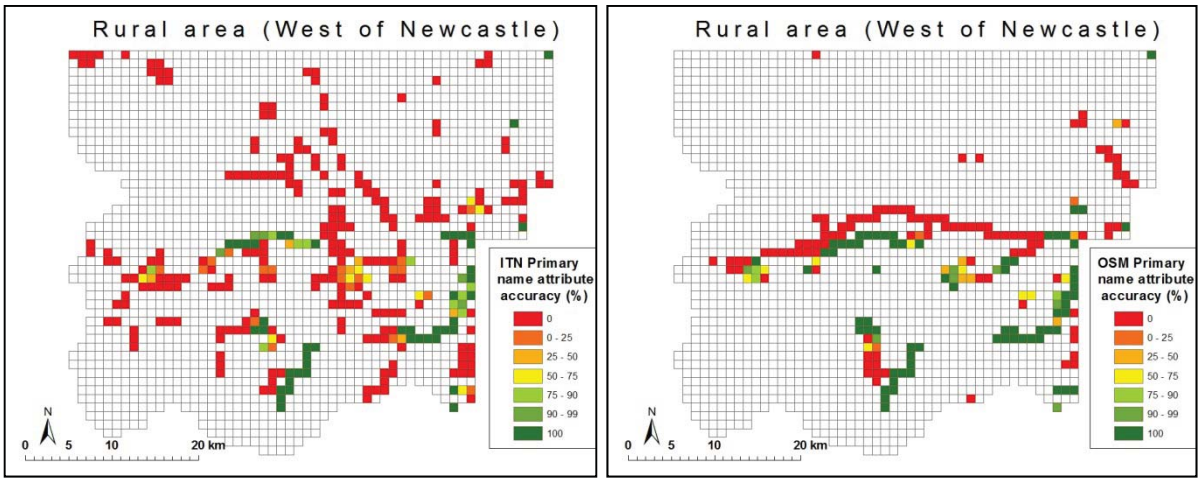

Figure 5.13: *Rural area – primary name,* ***a:*** *ITN percentages (OSM attribute accuracy)* ***b:*** *OSM percentages*

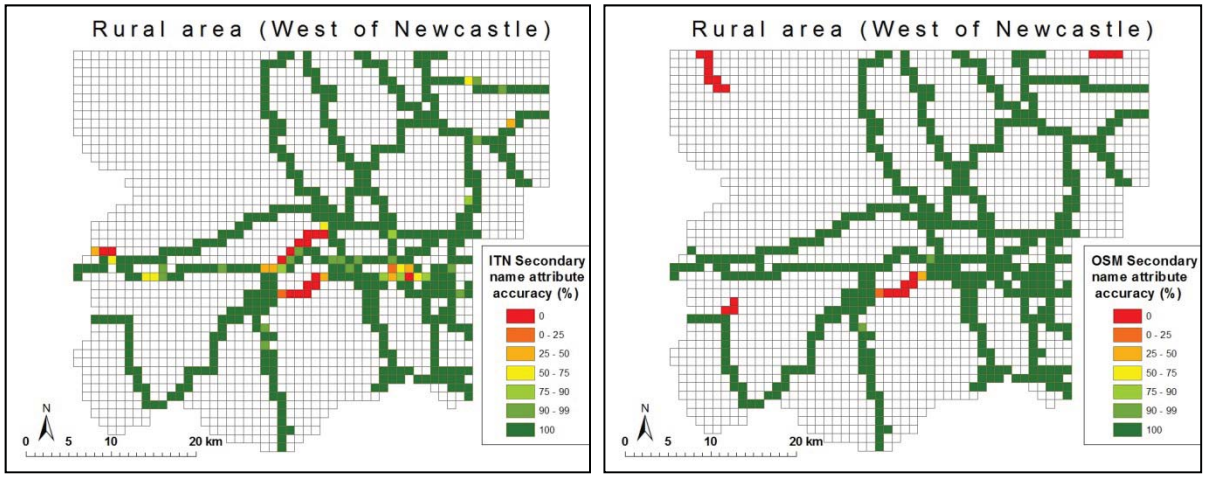

Figure 5.14: *Rural area – secondary name,* ***a:*** *ITN percentages (OSM attribute accuracy)* ***b:*** *OSM percentages*

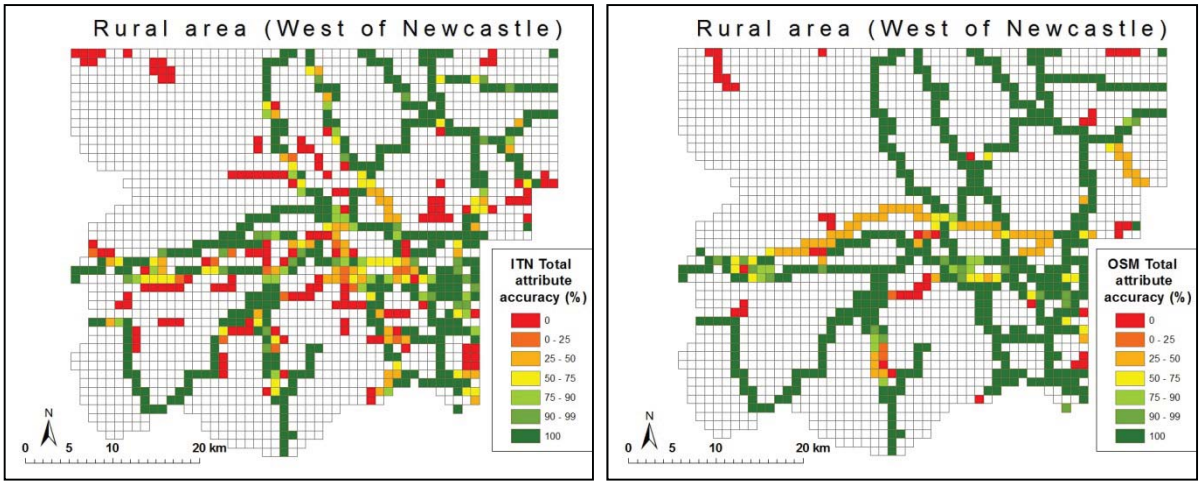

Figure 5.15: *Rural area – total names attribute accuracy,* ***a:*** *ITN percentages (OSM attribute accuracy)* ***b:*** *OSM percentages*

Similarly to the other quality elements, where there is no attribute for one dataset, the attribute accuracy value for the tile remains blank. This explains the 'linear' form of Figures 5.14a and 5.14b that examine the secondary name, which is used to define the rural network (e.g. A68, B6318, A696). All other tiles (white or blank) may have data with no such attribute. Accordingly, the same applies to the primary name (Figures 5.13a and 5.13b). Compared to the case of secondary names (Figures 5.14a and 5.14b), generally there can be information for the primary but not for the secondary name and vice versa. If there is information for at least one of the two names in one tile, it is further used when examining both names, otherwise tile quality value will be null for total attribute accuracy (Figure 5.15).

Zero values in dataset A and no values in dataset B means that there are name attributes present in dataset A for matched features that do not exist in dataset B, which also describes attribute commission (over-completeness). An example is the linear west to east road appearing in red in the middle of the tested rural area (Figure 5.13b, zero OSM attribute accuracy). This road has a secondary name 'B6318', correctly represented by both datasets (Figure 5.14). While in the ITN dataset it has no primary name (hence the empty values in Figure 5.13a), in the OSM dataset it has a primary name 'Military Road', which leads to zero percentage values in Figure 5.13b and affects OSM overall attribute accuracy (Figure 5.15b). Therefore, OSM has additional attribute information for this road, however whether this information is correct is a different matter. What is significant here is that the provided method and the selected classification of the results can spot such inconsistencies or differences.

Results' classification for data completeness and attribute accuracy is described in section 4.10 and 4.11.3 correspondingly. Figure 5.16 proves that data completeness is quite high, justifying the need to split the class 75%-99.99% in two, as section 4.10 described. For positional accuracy seven classes are applied. The first class refers to accuracies 0 to 8 m, which is the initial buffer used in this study and is considered a good quality for VGI. The seventh class includes the outliers (explained in section 4.12.4). The remaining ones are unevenly distributed to represent more appropriately the positional accuracy distribution of rural areas, which seems to be less homogeneous than in urban areas (Figures 5.8d and 5.12d).

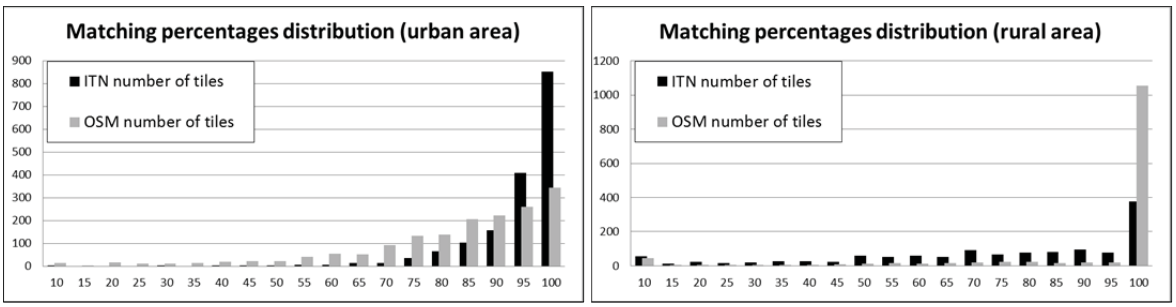

Figure 5.16: *Matching percentages' distribution for urban and rural areas*

Tables 5.3 and 5.4 provide more statistics based on the results for the tiles with the available data for each evaluation. Apart from the average value, values such as median, skewness, quartile and standard deviation show the distribution of the results, which is generally not normal. The meaning of these values is briefly presented for a better understanding of Tables 5.3 and 5.4.

Average is a value calculated by dividing the sum of values by the number of records (e.g. for values 0,0,0,4 the average is 1). Median is the actual value that separates the values in two (for the same example it is 0). Skewness shows the distribution trend: Negative skewness shows that the majority of values are concentrated on the right side of the distribution graph, while positive skewness shows the opposite (Figure 5.17, for the same example it is 2). Quartiles split values into four groups, and the 3rd quartile used here represents 75% of the results population (for the same example it is 1). Standard deviation, finally, shows the variation or dispersion of values compared to the average (for the same example it is 2).

Results are further discussed in section 5.5.

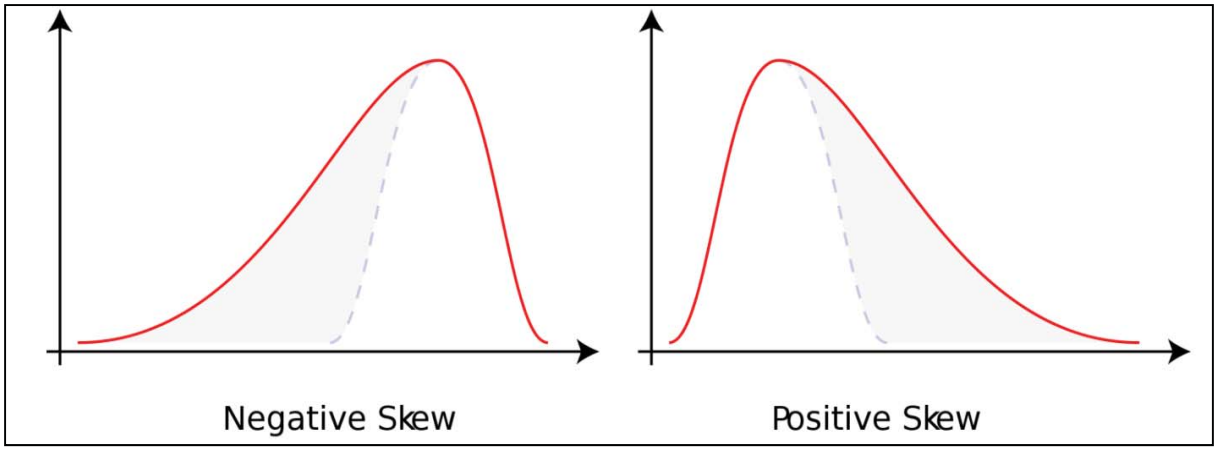

Figure 5.17: *Use of Skewness for data distribution*

Urban area (Greater London)	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	1,690	91.76	95.11	-3.85	98.33	11.23
	OSM	1,701	78.77	84.45	-1.67	93.26	20.53
Primary name accuracy	ITN	1,659	90.99	96.76	-3.66	99.05	17.50
	OSM	1,660	91.56	97.39	-3.80	99.99	17.36
Secondary name accuracy	ITN	1,372	90.31	99.05	-2.86	100.00	18.59
	OSM	1,387	93.32	100.00	-3.76	100.00	18.75
Total attribute accuracy	ITN	1,670	91.17	95.96	-3.94	98.69	15.09
	OSM	1,670	92.31	97.15	-4.11	99.76	14.68
Positional accuracy (outliers ignored)	OSM	1,634	11.38	10.50	4.92	11.75	4.59

Table 5.3: *Statistics for the urban area (Greater London)*

Rural area (W of Newcastle)	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	1,715	54.56	65.59	-0.33	91.93	39.27
	OSM	1,328	89.35	100.00	-2.61	100.00	25.84
Primary name accuracy	ITN	325	21.57	0.00	1.38	20.39	38.49
	OSM	201	43.10	0.00	0.27	100.00	47.28
Secondary name accuracy	ITN	488	94.62	100.00	-4.01	100.00	20.32
	OSM	494	95.45	100.00	-4.37	100.00	20.67
Total attribute accuracy	ITN	660	69.06	100.00	-0.85	100.00	41.18
	OSM	567	84.24	100.00	-1.84	100.00	30.67
Positional accuracy (outliers ignored)	OSM	1,181	15.56	10.28	1.45	22.50	13.26

Table 5.4: *Statistics for the rural area (West of Newcastle)*

5.4. Evaluation

5.4.1. Contribution of stages

Table 5.5 provides information on contribution of each stage to the matching procedure. Stage 1 that uses geometric only constraints and seeks for '1-1' matching, has more or less the same contribution regardless of the area type. Stages 2 and 3, which additionally examine thematic attributes, however, yield far more matches in urban areas. This is because VGI in rural areas is usually less complete in thematic attributes. Stage 4, on the contrary, which relies solely on geometry, is far more effective in rural areas, compensating for the less efficient stages 2 and 3. The reason for the low percentages in urban areas is that most of the data are already matched in the previous stages, so there are few objects left to be processed in this stage. Geometry of stage 4 is much more complex to handle when dealing with dense networks, so in this way data matching is more efficient, limiting the use of geometric-only constraints to cases with no other option due to the lack of thematic attributes. Stages 5 to 7 that deal with features aim to correct matching errors, so, depending on the data, they may increase or decrease the matching percentage achieved so far. 'Stage 5-' of Table 5.5 refers to the matched length re-classified as non-matched, so it is removed from the total matched length (hence the negative value). 'Stage 5+' is the opposite case. Further discussion regarding the sequence and contribution of the stages takes place in section 8.3.

Area	Dataset	Matched percentages (%) compared to the total matched length					
		Stage 1	Stage 2	Stage 3	Stage 4	Stage 5-	Stage 5+, 6, 7
Urban	ITN	22.40	**66.91**	1.81	8.88	-0.46	0.47
	OSM	31.01	**58.51**	1.67	6.76	-1.08	3.14
Rural	ITN	21.07	23.77	0.19	**56.09**	-2.43	1.31
	OSM	29.86	20.83	0.16	**47.97**	-1.05	2.23

Table 5.5: *Contribution of stages to data matching*

5.4.2. Object matching efficiency

A manual evaluation was performed for 10% of the tiles in both areas: 169 and 130 in urban and rural areas respectively. Tiles were randomly selected using a random number generation function, rejecting the ones with no data from both datasets. Figure 5.18 shows the evaluated tiles for both areas.

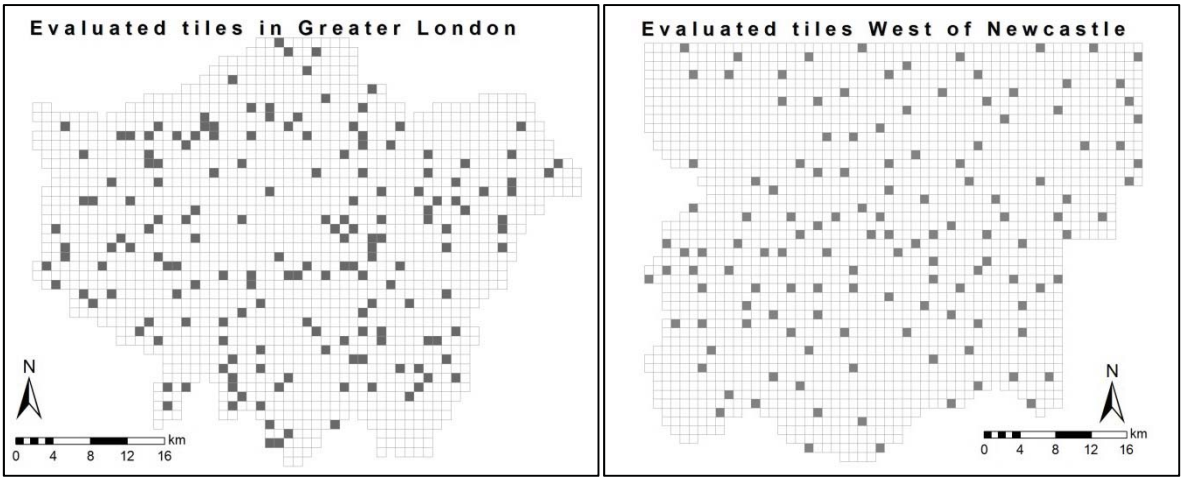

Figure 5.18: *Randomly selected tiles for manual evaluation*

The method is tested by calculating the length of the misjudged features for both datasets and comparing it with the dataset's length for each tile. Erroneous features are manually selected and marked, while their length is automatically calculated using an appropriate GIS software function. Figure 5.19 shows an example of the manual evaluation. Table 5.6 presents the results.

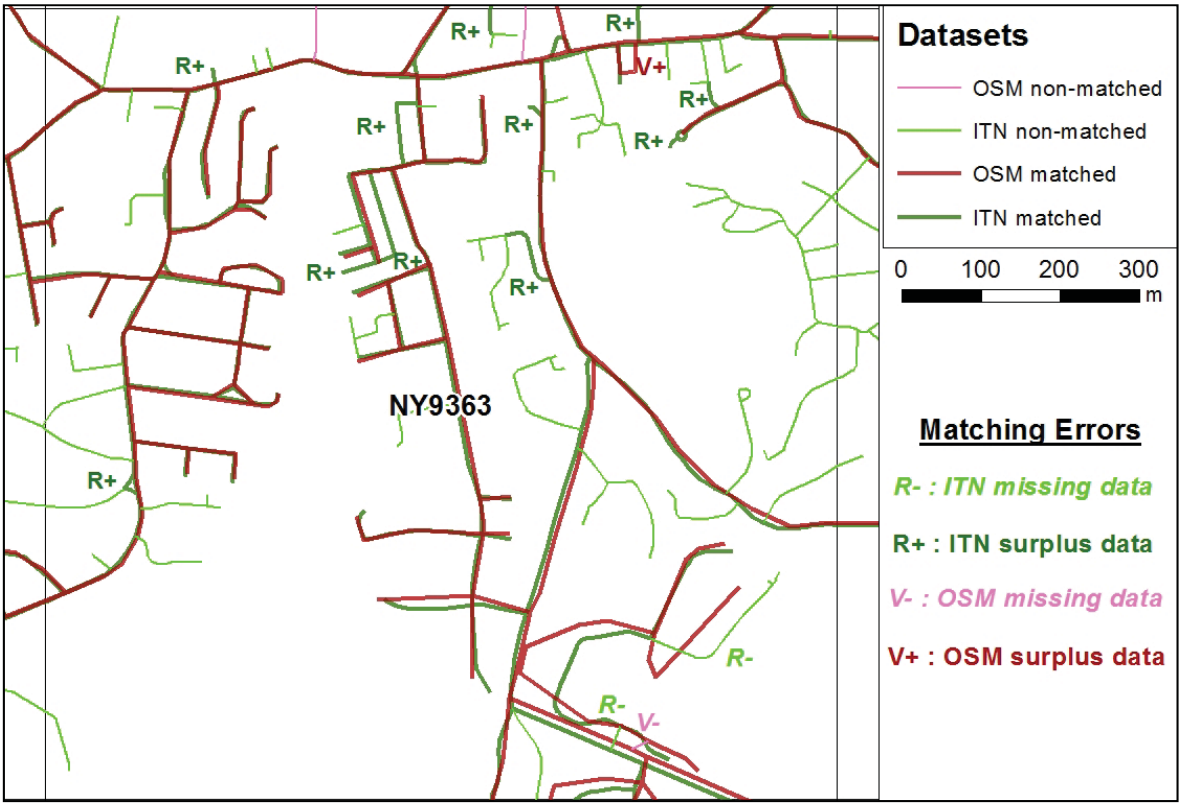

Figure 5.19: *Example of manually evaluating feature matching*

Area	*Data-set*	*Length (m) compared*	*Length (m) evaluated*	*Missing length (m)*	*Surplus length (m)*	*Total matching error (m)*
Urban	ITN	18,366,935	1,944,036(13.1%)	3,646(0.19%)	36,711(1.89%)	**40,357(2.08%)**
	OSM	20,719,274	2,202,873(13.0%)	13,004(0.59%)	14,147(0.64%)	**27,151(1.23%)**
Rural	ITN	2,500,826	287,942(11.5%)	4,098(1.42%)	5,643(1.96%)	**9,741(3.38%)**
	OSM	1,922,656	223,753(11.6%)	2,543(1.14%)	194(0.09%)	**2,737(1.22%)**

Table 5.6: *Data matching errors*

'Length compared' refers to the network length that the method was applied (explained in section 5.3), while 'Length evaluated' refers to the network length included in the tiles of Figure 5.18. In rural areas, errors are larger because the geometric constraints, although looser (see section 4.7), are sometimes too strict regarding the accuracy of data. On the VGI side, the satellite image used by OSM may be of lower resolution (e.g. Landsat). On the reference side, the ITN dataset may have a reduced accuracy (as section 5.1.1 discussed). This leads to corresponding roads being at distances even bigger than the search distances of looser constraints, e.g. 70 m or even more, and are classified as unmatched. Such rural inaccuracies, however, are sporadic and unpredictable, and may occur next to very accurate data (Figure 5.20). Using even looser constraints may favour these cases, but will not be suitable for more accurate data and will likely increase the amount of mistakenly matched objects in denser networks.

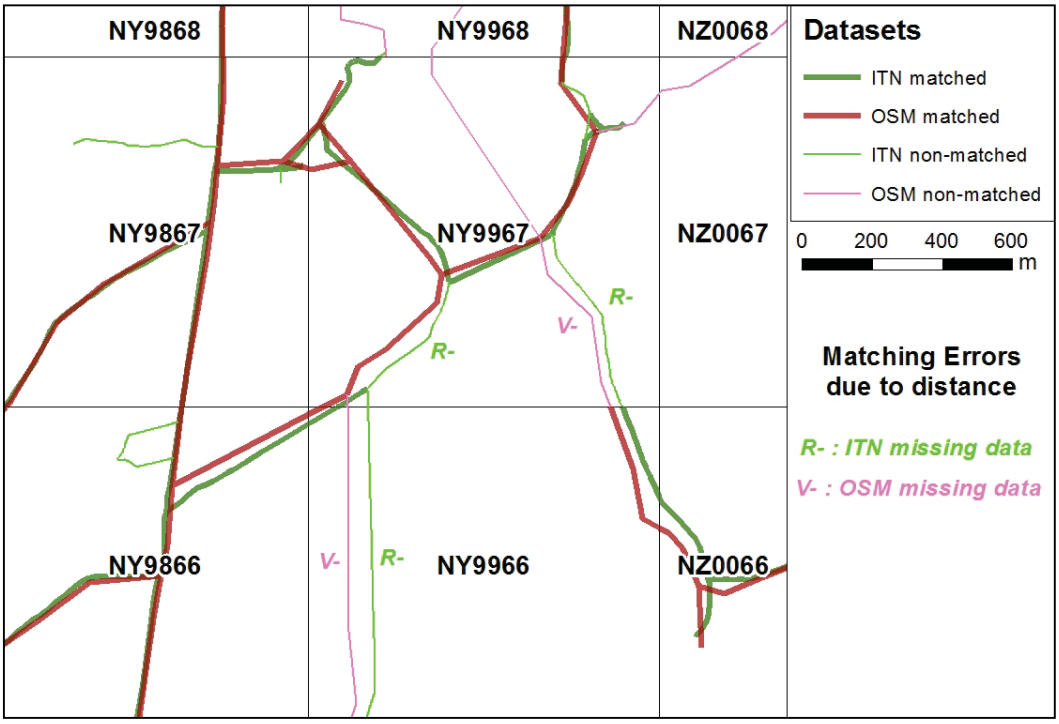

Figure 5.20: *Failure in matching corresponding objects in rural areas*

The amount of data manually examined for evaluation is more than adequate, as shown in Figure 5.21: error levels seem to remain the same when testing above 5% of randomly selected tiles.

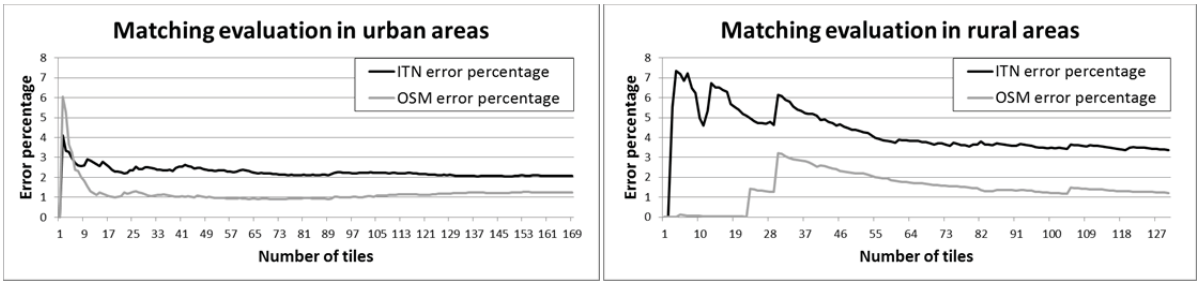

Figure 5.21: *Matching error compared to number of tiles evaluated*

5.4.3. Attribute accuracy efficiency

Section 4.11 described the method to measure attribute accuracy, while section 4.14 mentioned how this could be evaluated. This section provides a more detailed discussion on the three relevant types of error. Error type 1 refers to objects mistakenly considered as accurate (or non-accurate) in terms of attribute accuracy. This may happen when features are erroneously matched (data matching stage 5, section 4.8.7) or when attribute stage 3 fails by accepting a similar name which, although within the searching area mentioned in section 4.11.2, corresponds to a different feature. Error type 2 refers to failure of text similarity that results in accepting as accurate two different names (e.g. Richmond and Richland). Error type 3 refers also to failure in text similarity in the opposite way, which results in rejecting two similar names.

Figure 5.22 provides an example of the manual evaluation of attribute accuracy, showing some of the above cases, as well as how attribute accuracy is calculated. Primary and secondary names are separated by an asterisk. Numbers in parentheses refer to the stage of attribute matching (1 for exact name matching, 2 for similar name and 3 for similar name within a distance), while zero means that there is no match found for this name. Green lines and capital letters refer to the ITN dataset, while red lines and small letters refer to OSM.

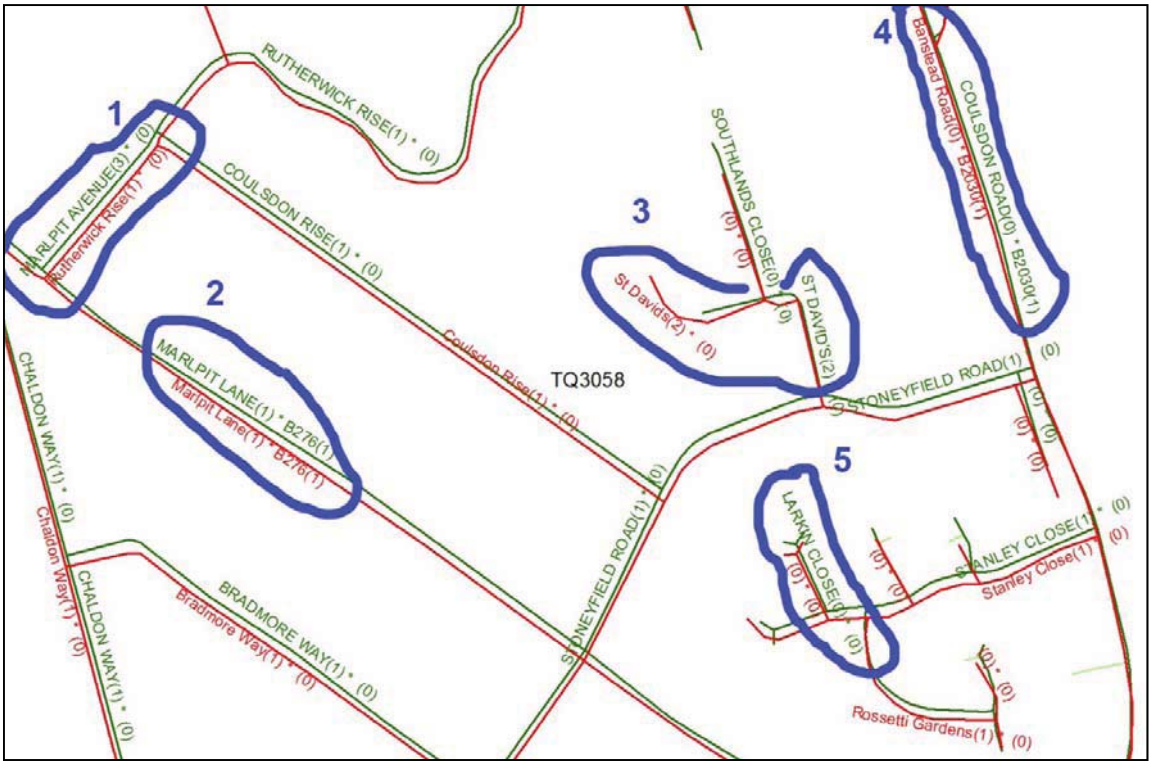

Figure 5.22: *Manual evaluation of attribute accuracy measurement*

In case 1 there is an error type 1, accepting corresponding objects named 'Marplit Avenue' and 'Rutherwick Rise' as accurate in the 3rd and 1st stage respectively. In case 2 there is an example of a feature with both names correct, while in case 4 only the secondary name is correct. Case 3 is an example of the text similarity efficiency. In case 5, finally, there is no match for the ITN name, which shows the reference dataset superiority regarding thematic attributes.

For the urban area, 55 tiles were randomly selected from the ones used in data matching evaluation (Figure 5.18a). For the rural area, however, attribute information scarcity enabled the evaluation of the total number of the tested tiles of Figure 5.18b. Total results are presented in Tables 5.7 and 5.8 for the urban and rural areas respectively. Errors refer to the feature level and the length is used to estimate the total error for each dataset as a percentage value. For the erroneous features found, their length is compared with the sum of features with primary and features with secondary name. As a result, the length of every feature with both names present will be calculated twice. This is to cover cases where a feature has both names but only one of them is correct or present in one dataset.

The low total error percentages for the attribute accuracy method (1.47% and 1.41% for urban and rural areas respectively) show the efficiency of the method.

Urban area (Greater London)	ITN matched dataset		OSM matched dataset	
	Length (m)	Pct (%)	Length (m)	Pct (%)
Total attributes	520,312.6	100.00	487,706.6	100.00
Primary name	429,008.3	82.45	407,166.1	83.49
Secondary name	91,304.3	17.55	80,540.5	16.51
Error type 1	5,919.9	1.14	3,258.3	0.67
Error type 2	383.9	0.07	418.4	0.09
Error type 3	1,350.0	0.26	1,823.0	0.37
Total errors	7,653.8	**1.47**	5,499.7	**1.13**

Table 5.7: *Attribute accuracy errors for the urban area*

Rural area (West of Newcastle)	ITN matched dataset		OSM matched dataset	
	Length (m)	Pct (%)	Length (m)	Pct (%)
Total attributes	91,399.1	100.00	82,711.0	100.00
Primary name	33,971.5	37.17	27,583.6	33.35
Secondary name	57,427.6	62.83	55,127.4	66.65
Error type 1	560.5	0.61	12.8	0.02
Error type 2	0.0	0.00	0.0	0.00
Error type 3	729.8	0.80	681.0	0.82
Total errors	1,290.3	**1.41**	693.8	**0.84**

Table 5.8: *Attribute accuracy errors for the rural area*

5.4.4. Positional accuracy efficiency

Positional accuracy, as was described in section 4.12, is calculated using increasing buffers. Generally, positional accuracy measurement may be affected in two ways. Errors in data matching, specifically mistakenly accepted-as-matched VGI features, can only be intersected when using a bigger buffer size on reference objects that are relatively far away and obviously non-corresponding (Figure 5.23a). The same happens when the VGI dataset contains longer features to represent the same road network object (Figure 5.23b). Both cases lead to lower positional accuracy values (bigger buffer sizes) that may not be so representative. What seems to be supportive to the provided method, however, is that the opposite case is unlikely to happen: The above cases of erroneous data matching and different representation on reference instead of VGI dataset will not lead to a higher accuracy value, because the additional buffer will not include any additional VGI dataset (e.g. Figure 5.24 - upper left part with the erroneously matched ITN feature and no corresponding OSM one).

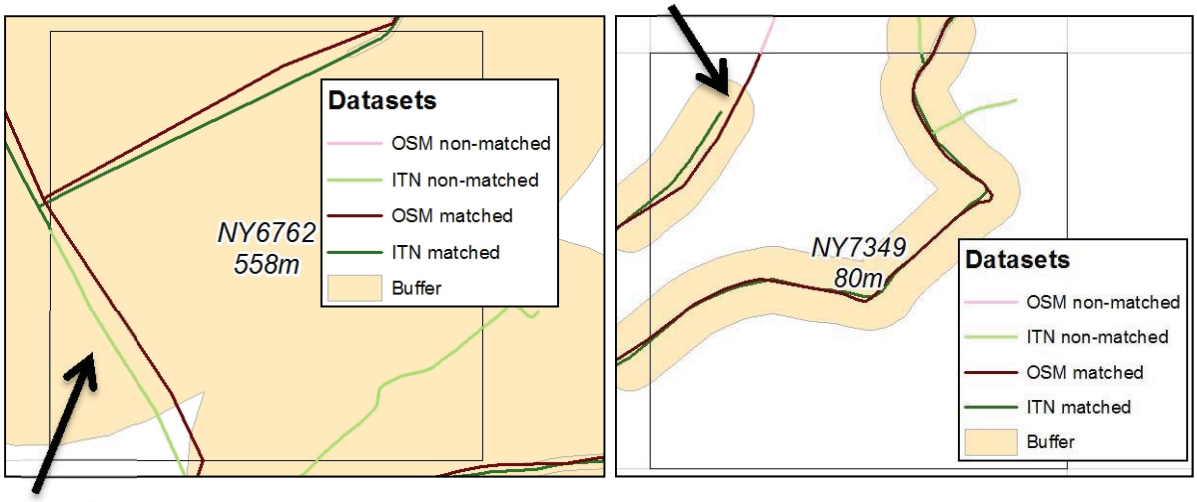

Figure 5.23: *Positional accuracy and* ***a:*** *Data matching errors,* ***b:*** *Different representation*

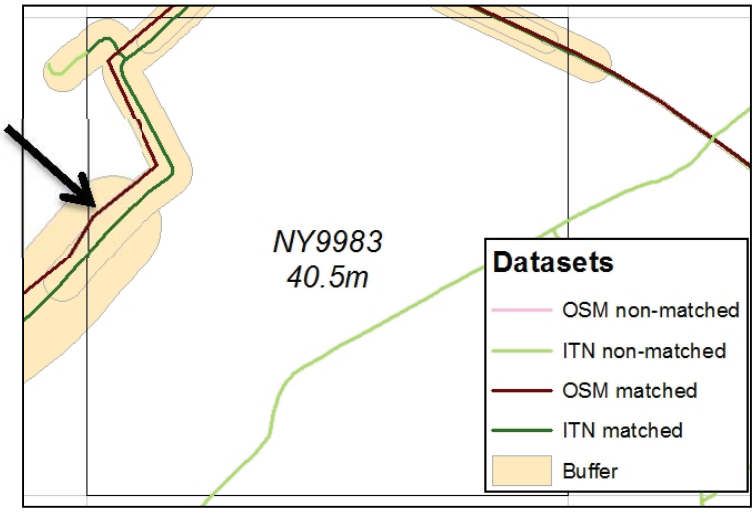

Figure 5.24: *Positional accuracy and big distance between corresponding objects*

However, abnormally high buffer values are treated as outliers, as explained in section 4.12.4. For this study case, all outliers in urban and rural area (Figures 5.8d and 5.12d respectively) are visually examined to justify their existence. There are three reasons found:

- errors in data matching (e.g. Figure 5.23a),
- increased distances between corresponding objects (e.g. Figure 5.24), and
- different representations for the same object. Figure 5.23b provides a simple case where the OSM object is slightly extended, however there are cases where the extension can be quite significant, or objects can even change direction and shape. In such cases of partial matching it is difficult to decide for the total feature even if data matching was manually performed.

Table 5.9 provides the results of the manual evaluation. The percentage of total outliers is calculated with respect to the total tiles compared, while the percentages for each of the three reasons refer to the total number of outliers.

	Urban Area		Rural Area	
	Number	**Percentage**	**Number**	**Percentage**
Total outliers	47	2.79% of total tiles	72	5.67% of total tiles
Data matching errors	19	40.43%	42	58.33%
Distance errors	2	4.26%	3	4.17%
Different representation	26	55.32%	27	37.50%

Table 5.9: *Evaluation of positional accuracy outliers*

Outliers are far more common in the rural area (5.67% compared to 2.79% in the urban area). 'Different representation' is the most significant factor in urban areas, while in rural areas 'Data matching errors' prevails. However, it was noticed that the majority of outlier tiles with data matching errors in rural area is attributed to the increased distance between some corresponding objects, according to which the OSM feature is accepted as matched, but its corresponding ITN feature is not matched. 'Distance errors' factor refers to data correctly matched but with an increased distance between them. Such a case should not be considered as an outlier, since the increased buffer value represents correctly the reduced tile positional accuracy. The relatively low number of outliers attributed to 'Distance errors' suggests that the thresholds used to classify a tile as an outlier are efficiently selected for both urban and rural cases.

5.4.5. VGI commission indication

As sections 4.13.2 and 4.13.3 explained, some non-matched VGI features are marked as possible new features, based on their attributes or road type. This, however, needs to be evaluated in order to see the extent of data matching errors that it may include, which would result in objects that should have been matched instead of indicated as unique or commissioned ones. The same tiles of Figure 5.18 were manually evaluated and results (Table 5.10) show error levels of 3.42% and 17.44% respectively for each case. The second one is higher, reflecting the bigger inconsistencies between the examined datasets that lead to higher data matching errors in rural areas, as also shown in Table 5.6.

Evaluation of VGI Commission indication	Urban OSM dataset		Rural OSM dataset	
	Number	Length(m)	Number	Length(m)
All non-matched objects	5,703	509,549 (100%)	109	32,014,603 (100%)
Indicated non-matched objects as new	1,920	143,812 (28.22%)	79	22,186 (69.3%)
Data matching errors (instead of new objects)	55	4,911 **(3.42%)**	7	3,870 **(17.44%)**

Table 5.10: *Evaluation of VGI commission indication*

Noteworthy, however, this manual evaluation refers to whether a feature should be considered as new (non-existing in the other dataset) or not, regardless of its road type. In other words, indicated objects as new ones can also be footpaths or other data types not present in the reference dataset. The reason is that the term 'Commissioned' data (mentioned in section 4.13) depends also on what this method is used for. If for example we need to enhance the reference dataset according to its specifications, non-matched features that are not needed will be removed (e.g. footpaths, steps, cycleways in this case). If, on the other hand, a different dataset with additional types of information is needed (e.g. a dataset for pedestrians as well), the appropriate new features should be selected accordingly. As a result, marking a VGI feature as 'new' should cover all possible scenarios. However, due to the fact that this marking includes also data matching errors, a manual evaluation is necessary if the non-matched VGI features are to be further used for conflation purposes.

5.5. Discussion

5.5.1. Data matching errors and quality results

As described in Chapter 4, data matching is the first step before any further quality analysis, so data matching errors will affect data completeness, attribute and positional accuracy. Using the equation 11 of section 4.14 and the results of Table 5.6, VGI completeness results from Table 5.2 (matched length) are 0.54% to 1.70% higher than they should be, for urban and rural areas respectively.

As noted in section 4.14, it is difficult to calculate how attribute accuracy is affected by data matching errors, however it is anticipated to be less significant than their impact on data completeness. The evaluation results of section 5.4.3 seem to agree with this assumption. Using equation 12 of section 4.14, OSM attribute accuracy values (as shown by ITN percentages in Tables 5.3 and 5.4) are estimated to differ less than 1% from the actual values.

Positional accuracy calculation, finally, is affected by data matching errors only when they refer to mistakenly matched VGI or/and rejected reference features. In these cases this leads to a lower accuracy value than the actual one, however it is difficult to predict when and where this will occur. Combined with other factors, such as the desired overlap percentage, the bigger distances between corresponding objects and different representation, positional accuracy can be so low for some tiles (meaning high buffer value), that they will be rejected as outliers. Using the results of Table 5.9 and equation 14 of section 4.14, the percentage of outliers caused by data matching errors is 1.13% and 3.31% in the urban and rural area respectively.

Considering that generally the urban or rural nature of the road network will not prevail to the extent of the selected areas of this chapter, it is assumed that for an area of mixed network density error levels will be between the values calculated here for each area type. Table 5.11 provides a generalised estimation of the quality errors for the provided methodology (regardless of the area classification as urban or rural), based on the manual evaluation results described in section 5.4.

VGI spatial quality element	Estimated error range
OSM Completeness (based on data matching)	From +0.54% to +1.70%
Attribute accuracy	From - 0.19% to +0.95%
Positional accuracy	From 1.13% to 3.31% of outliers

Table 5.11: *Estimation of errors in quality results for the provided method*

5.5.2. Road types correspondence

Section 4.10 described how road type correspondence information is collected. As shown in Table 4.3, correspondence for each road type can be with many from the other dataset, however not all of them represent actual correspondence, either because of data matching errors, or due to the different classification between the datasets, or even because of errors in VGI classification. The first one (or two, depending on the percentage) corresponding types are considered as most important and are presented in Tables 5.12 and 5.13. While for some road types their correspondence is obvious regardless of the area (e.g. 'B Road' with 'secondary'), for others it is more difficult when first and second percentages are close, or when they differ significantly between urban and rural areas.

Urban Area			Rural Area		
ITN Road Type	**OSM Road type**	**%**	**ITN Road Type**	**OSM Road type**	**%**
A Road	primary	58.9	**A Road**	trunk	67.3
	trunk	36.0		primary	32.0
Alley	residential	43.5	**Alley**	residential	63.8
	service	33.2		service	14.1
B Road	secondary	90.9	**B Road**	secondary	96.0
	residential	3.1		unclassified	2.5
Local Street	residential	85.8	**Local Street**	residential	38.7
	unclassified	9.1		unclassified	23.6
Minor Road	residential	39.0	**Minor Road**	unclassified	52.1
	tertiary	36.3		road	25.2
Motorway	motorway	87.7	**Motorway**	(road type not present)	
	motorway_link	11.6			
Pedestrianised Street	pedestrian	56.5	**Pedestrianised Street**	(road type not present)	
	unclassified	14.2			
Private Road - Publicly Accessible	service	56.7	**Private Road - Publicly Accessible**	footway	27.6
	residential	16.4		road	27.2
Private Road - Restricted Access	residential	35.2	**Private Road - Restricted Access**	track	29.6
	service	31.2		unclassified	21.2

Table 5.12: *ITN road types correspondence*

By combining Tables 5.12 and 5.13, corresponding road types bilaterally agreed regardless of the area are:

- ITN Motorway with OSM motorway,
- ITN A Road with OSM primary or trunk,
- ITN B Road with OSM secondary, and
- ITN Local Street with OSM residential.

For all other road types, the link is neither two-way between datasets, nor the same for urban and rural areas. This information helps deciding on the importance of non-matched OSM features during their examination regarding VGI commission. So, it is more important for example to examine non-matched OSM features with ‘secondary’ road type than ‘cycleway’ or ‘footway’.

Urban Area			Rural Area		
OSM Road Type	**ITN Road type**	**%**	**OSM Road Type**	**ITN Road type**	**%**
access	Private Road - RA*[6]	82.8	**access**	Private Road - RA*	73.0
	Local Street	15.5		Local Street	27.0
bridleway	Private Road - RA*	53.5	**bridleway**	Private Road - RA*	91.8
	Local Street	42.5		Local Street	7.1
cycleway	Local Street	38.5	**cycleway**	A Road	57.1
	A Road	23.6		Local Street	17.8
footway	Private Road - RA*	44.7	**footway**	Local Street	44.5
	Local Street	35.3		Private Road - RA*	28.1
living_street	Local Street	61.9	**living_street**	(road type not present)	
	Private Road - RA*	37.9			
motorway	Motorway	97.7	**motorway**	(road type not present)	
	Private Road - RA*	1.3			
motorway_link	Motorway	78.0	**motorway_link**	(road type not present)	
	Local Street	13.3			
path	Private Road - RA*	47.8	**path**	Private Road - RA*	76.6
	Alley	16.1		Minor Road	13.9
pedestrian	Local Street	40.0	**pedestrian**	Private Road - RA*	58.3
	Private Road - RA*	28.4		Local Street	41.7
primary	A Road	92.5	**primary**	A Road	93.5
	Local Street	3.6		Private Road - RA*	3.1
primary_link	A Road	77.1	**primary_link**	(road type not present)	
	Local Street	8.8			
residential	Local Street	87.7	**residential**	Local Street	79.0
	Minor Road	6.8		Private Road - RA*	9.9
road	Local Street	52.9	**road**	Minor Road	81.1
	Private Road - RA*	30.6		Private Road - RA*	9.7
secondary	B Road	86.6	**secondary**	B Road	94.6
	Minor Road	5.9		Private Road - RA*	3.2
service	Private Road - RA*	45.4	**service**	Private Road - RA*	63.7
	Local Street	31.2		Local Street	21.5
tertiary	Minor Road	82.6	**tertiary**	Minor Road	91.8
	Local Street	13.3		Private Road - RA*	3.9
track	Alley	43.0	**track**	Private Road - RA*	62.1
	Private Road - RA*	40.3		Minor Road	31.6
trunk	A Road	93.4	**trunk**	A Road	92.2
	Local Street	3.8		Private Road - RA*	4.4
trunk_link	A Road	66.1	**trunk_link**	A Road	84.6
	Minor Road	17.7		Private Road - RA*	12.7
unclassified	Local Street	61.8	**unclassified**	Minor Road	87.6
	Minor Road	23.4		Private Road - RA*	6.1

Table 5.13: *OSM road types correspondence*

[6] RA*: Restricted Access

On the other hand, the difficulty in matching road types shows semantic fuzziness between the different classifications. This is the result of having different classes with different definitions between datasets. The problem is also mentioned by Kounadi (2009) and Girres and Touya (2010). Additionally, there are cases of classification errors of VGI contributors, which are generally easy to spot: the data amount (in terms of length) for the misclassified pair is insignificant compared to the other pairs for the same class. Table 4.3 demonstrates an example: OSM features that were matched to ITN's 'B Road' would normally not be 'private', 'path', 'pedestrian', 'steps', 'footway'. Additionally, ITN's 'Pedestrianised Street' or 'Alley' would not be OSM's 'secondary' road type. However, although errors in classification are easy to spot for percentages close to zero, the line between user error and semantic fuzziness is also fuzzy. For example, it is difficult to decide where ITN's 'B Road' cases linked to OSM's 'tertiary' or 'trunk' belong.

5.5.3. VGI commissioned data

Road type correspondence can further assist in finding commissioned data. By calculating the amount of each road type that is found with a match in the other dataset, valuable information can be collected on what is generally mapped or failed to be mapped. Table 5.14 refers to the ITN standardised road types in the urban and rural areas tested. OSM seems failing to map the 'Alley' road type in both urban and rural areas. The next 'neglected' ITN road type is 'Private Road – Restricted Access', which although in the urban area is not as bad, in the rural one it reaches 88%. ITN supremacy in urban areas refers mainly to these two road types, while results in the rural area show again that generally VGI is by far less complete.

ITN road type	Urban area		Rural area	
	Matched length (m)	Non-matched length (m)	Matched length (m)	Non-matched length (m)
A Road	2,369,890 (99.96%)	1,007 (0.04%)	218,134 (100%)	0 (0%)
Alley	289,244 (30.77%)	650,828 **(69.23%)**	1,879 (42.78%)	2,513 **(57.22%)**
B Road	539,875 (99.98%)	116 (0.02%)	255,078 (99.63%)	937 (0.37%)
Local Street	10,684,372(98.38%)	176,181 (1.62%)	127,180 (44.95%)	155,773 (55.05%)
Minor Road	1,841,015 (99.93%)	1,252 (0.07%)	1,010,315 (87.90%)	139,123 (12.10%)
Motorway	137,887 (100%)	0 (0%)	(road type not present)	
Pedestrianised Street	12,812 (98.16%)	240 (1.84%)	(road type not present)	
Private Road - Publicly Accessible	120,738 (76.47%)	37,146 (23.53%)	2,395 (36.60%)	4,149 (63.40%)
Private Road - Restricted Access	1,088,706 (72.31%)	416,841 **(27.69%)**	120,714 (11.86%)	897,482 **(88.14%)**

Table 5.14: *ITN matched road types: what is mapped by OSM and what is not*

OSM road type	Urban area (m & percentage)		Rural area (m & percentage)	
	Matched length	Non-matched length	Matched length	Non-matched length
abandoned	613 (100%)	0 (0%)	(road type not present)	
access	2,342 (40.4%)	3,456 (59.6%)	74 (100%)	0 (0%)
bridleway	25,986 (23.71%)	83,605 (76.29%)	7,213 (26.28%)	20,237 (73.72%)
byway	73 (2.52%)	2,805 (97.48%)	(road type not present)	
construction	762 (56.31%)	592 (43.69%)	(road type not present)	
conveyor	0 (0%)	119 (100%)	(road type not present)	
crossing	26 (50.28%)	26 (49.72%)	(road type not present)	
cycleway	12,798 (3.75%)	328,233 (96.25%)	0 (0%)	15,296 (100%)
depot	0 (0%)	813 (100%)	(road type not present)	
Foot Path	0 (0%)	118 (100%)	(road type not present)	
footway	213,779 (8.23%)	2382,780 (91.77%)	4,976 (4.98%)	94,933 (95.02%)
footway unoffic	0 (0%)	242 (100%)	(road type not present)	
footway; pedestr	0 (0%)	485 (100%)	(road type not present)	
ford	(road type not present)		22 (100%)	0 (0%)
living_street	5,269 (97.89%)	114 (2.11%)	(road type not present)	
motorway	113,544 (100%)	0 (0%)	(road type not present)	
motorway_link	33,269 (93.39%)	2,356 (6.61%)	(road type not present)	
path	6,583 (3.87%)	163,622 (96.13%)	5,681 (16.81%)	28,104 (83.19%)
pedestrian	38,053 (37.25%)	64,097 (62.75%)	0 (0%)	463 (100%)
platform	0 (0%)	592 (100%)	(road type not present)	
primary	1363,227 (99.91%)	1,167 (0.09%)	69,616 (99.86%)	98 (0.14%)
primary_link	19,845 (95.78%)	874 (4.22%)	(road type not present)	
private	97 (16.08%)	504 (83.92%)	(road type not present)	
proposed	0 (0%)	177 (100%)	(road type not present)	
res	63 (100%)	0 (0%)	(road type not present)	
residential	10297,620(98.97%)	107,296 (1.03%)	60,129 (98.49%)	919 (1.51%)
residential; unc	724 (100%)	0 (0%)	(road type not present)	
residential;uncl	325 (100%)	0 (0%)	(road type not present)	
road	19,118 (59.67%)	12,924 (40.33%)	312,655 (90.35%)	33,380 (9.65%)
secondary	525,606 (99.88%)	646 (0.12%)	247,237 (100%)	0 (0%)
secondary_link	185 (100%)	0 (0%)	(road type not present)	
service	762,671 (53.81%)	654,610 (46.19%)	12,469 (72.24%)	4,791 (27.76%)
services	3,801 (100%)	0 (0%)	(road type not present)	
steps	962 (4.73%)	19,384 (95.27%)	0 (0%)	198 (100%)
tertiary	755,042 (99.82%)	1,370 (0.18%)	205,254 (98.86%)	2,357 (1.14%)
tertiary_link	168 (79.92%)	42 (20.08%)	(road type not present)	
track	66,103 (40.92%)	95,437 (59.08%)	65,617 (78.71%)	17,744 (21.29%)
trunk	797,511 (99.9%)	779 (0.1%)	146,346 (100%)	0 (0%)
trunk_link	91,899 (95.88%)	3,951 (4.12%)	2,197 (100%)	0 (0%)
unclassified	1557,799 (95.14%)	79,514 (4.86%)	579,948 (97.53%)	14,673 (2.47%)
unsurfaced	2,859 (52.1%)	2,629 (47.9%)	(road type not present)	

Table 5.15: *OSM matched road types: what is mapped by ITN and what is not*

On the other side of the comparison, Table 5.15 refers to the OSM road types in the urban and rural areas tested. Not all road types are present in both areas for both datasets. The increased number of OSM road types makes it less easy to interpret Table 5.15, however there are three major conclusions.

- The increased number of road types and their inconsistency between the two areas is a result of no standards in OSM. Although there is a suggestion on road types tagging, it seems that it is not always followed, which leads to customised road types that complicate the network classification. This is also another consequence of VGI heterogeneity.
- The fact that even unusual road types (e.g. 'ford', 'res', 'services') may be 100% matched, which means that they are also present in the ITN dataset, makes it difficult to decide on a connection between specific ITN and OSM road types, so that by rejecting these OSM road types the datasets would be more homogeneous in terms of the objects represented. Even OSM road types that are not traditionally mapped by ITN (e.g. cycleways, footpaths, bridleways, paths) were partially matched, at such percentages that it cannot be the result of erroneous data matching, considering the efficiency of the method. This is mostly a result of erroneous road classification in OSM, which cannot be otherwise tested since the classification criteria differ between the datasets. It makes it obvious, however, that all VGI features need to be compared, otherwise by removing selected road types, corresponding data may be rejected, which will give a false estimation of data completeness, attribute and positional accuracy.
- Adding to the previous conclusion, the same road type may have a totally different behaviour between different areas, e.g. 'tracks' have more corresponding ITN data in rural than in urban areas (79% and 41% respectively), so it is even harder to decide on what to reject before starting the evaluation process.

Similarly to the ITN road types, OSM road types with a low matched percentage refer to data generally not mapped in ITN, so they show where OSM contains additional information, as compared to the ITN dataset, and can be used to narrow down OSM data selection when needed for conflation purposes.

According to section 5.4.5, OSM non-matched features marked as possible commissioned data were manually examined. In many cases the findings of this approach were matching errors, making it an efficient quality measure (Figure 5.25a). In few cases it also showed VGI inconsistencies (Figure 5.25b). However, there are also cases where VGI superiority towards the reference dataset is

successfully found. Such a manual examination can be easier when using satellite imagery as a background (assuming the image is correctly georeferenced). This is possible through ArcGIS software in two ways: Firstly, ArcMap Version 10 allows the use of base maps, among which are 'Bing Maps' with Yahoo satellite imagery. Secondly, non-matched VGI shapefiles can be exported in KML format, along with the initial reference dataset, so that they could be loaded on Google Earth software. In both cases, however, output datasets have to be reprojected to WGS84, which is the reference system of both base maps.

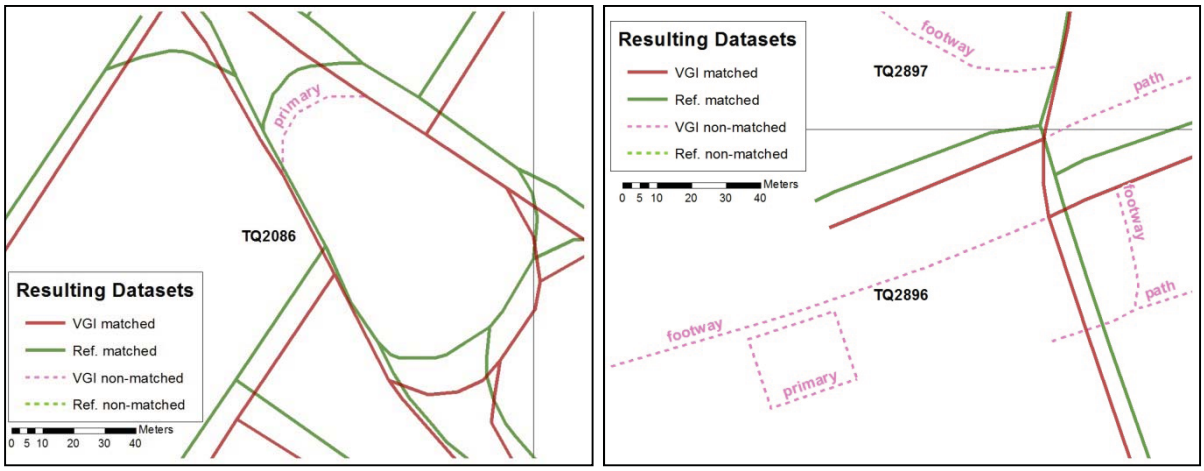

Figure 5.25a: *Data matching error (Primary road corresponding to ITN's A Road, not matched),* ***b:*** *OSM inconsistency: Rectangle defined as primary road*

Two examples of the second option are in Figures 5.26 and 5.27, where yellow lines represent the reference dataset (as a whole) and red ones the OSM non-matched dataset. These are roads that could be used by a normal family car, so according to the ITN specifications they should have been present in the ITN dataset too. Additional examples can be found in Appendix C.

5.5.4. Topology

Section 3.2 discussed the broader definition of topology. In this context, however, it is limited to describe the fact that objects are represented appropriately as polylines, by being segmented at road junctions or when attributes change (e.g. a straight road that changes name in the middle). There are also cases where a junction should not be placed in a road intersection and intersected lines need not be divided. Such is the case of overlapping roads through bridges, where although polylines intersect, they do not form crossroads. More advanced topological requirements include direction of the objects, which is necessary when data are used for routing purposes, and end-points

matching, which means that two adjacent linear objects should have one common end-point. These, however, are not examined here.

Figure 5.26: *Rural area, OSM commission*

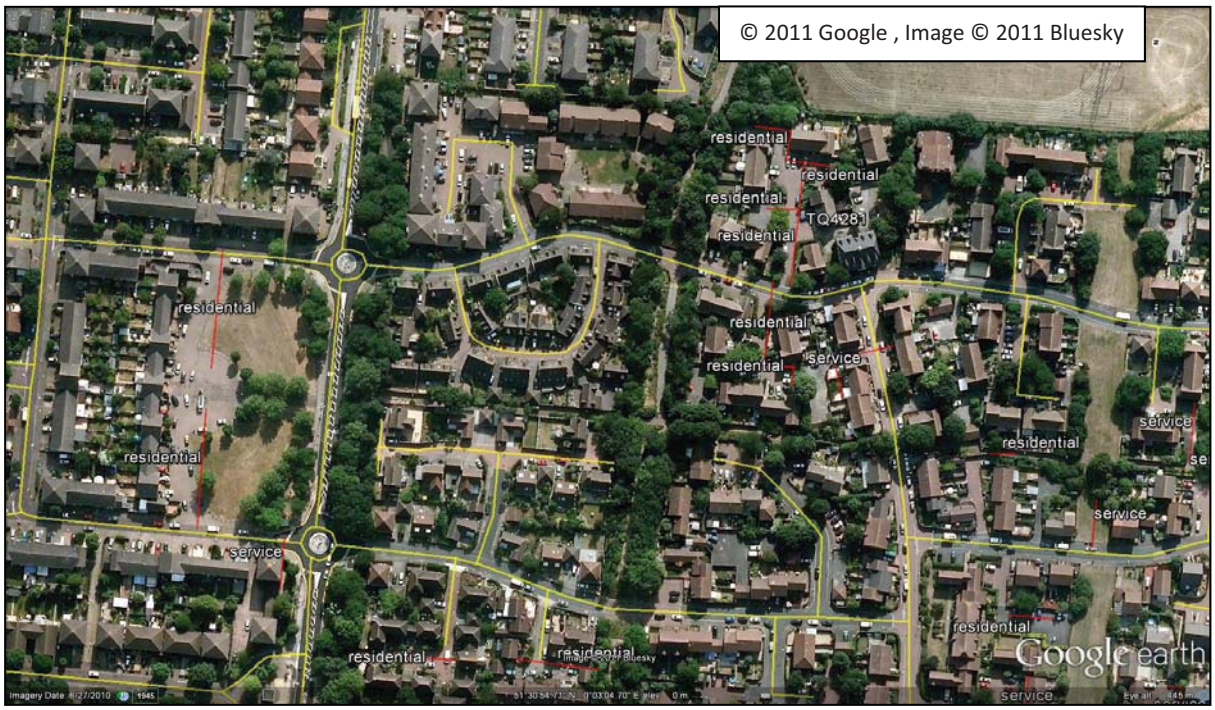

Figure 5.27: *Urban area, OSM commission (image rotated), tile TQ4281*

Topology of ITN is expected to be error-free, since the dataset is designed for routing purposes. Indeed, no such errors were identified during the manual evaluation of data matching. OSM on the other hand does not have a similar objective or a standardised procedure that will rule out the possibility of topological errors. The manual data matching evaluation showed that OSM has generally a good topology (though not without errors), which was noticed to be better in urban and worse in rural areas. The question is whether an automated creation of topology by splitting features when intersected would improve data matching, considering that in some cases features need to be split and in others not. If, for example, an OSM feature is extended beyond a road junction by including a road object not present in ITN dataset, it should be partially matched, since there is a corresponding object for part of it. However, the data matching examination classifies objects on a feature level and not partially, so depending on the constraints used, this feature will be classified as a whole either as matched or as non-matched (an example of the latter was provided in Figure 5.20 for the non-matched OSM and ITN features of tile NY9967). By correcting the topology, this feature would be divided at the road junction and only one of the new features would be efficiently matched with its corresponding object. On the other hand, however, for cases where there should not be any intersection (e.g. bridges), the automated topology correction would split the OSM feature and its derivatives would have to be compared with a longer ITN feature, which is not split accordingly, so data matching may be negatively affected.

To test this question, the topology of the rural OSM dataset was corrected accordingly, splitting features when intersected and creating new ones, and the data matching process was repeated. The rural dataset was selected because of the increased number of topological errors found during the manual evaluation of data matching. Results are presented in Table 5.16. When compared with the results of Table 5.2, differences prove to be insignificant: only 68 m (0.01%) added to the OSM matched dataset, while all other lengths remain unchanged. This shows that topology does not have to be examined or corrected in this case study. However, the use of different sources will test if the method is generally robust and efficient regardless of the topology.

Area	Dataset	Total Length	Length Compared	Length Matched	Length non-matched
Rural – topologically corrected	ITN	2,935,675 m	2,500,826 (85.19 %)	1,735,695 (59.12 %)	1,199,980 (40.88%)
	OSM	1,952,443 m	1,922,468 (98.46 %)	1,719,503 (88.07 %)	232,940 (11.93%)

Table 5.16: *Results for rural area with corrected topology*

5.5.5. Spatial Patterns

To conclude the discussion, some additional spatial patterns and findings regarding Figures 5.8 to 5.15 need to be mentioned. There are some edge effects around London (Figures 5.8 to 5.11). Although both datasets were clipped similarly, there are boundary tiles where OSM dataset seems richer. This is described by tiles that are blank in Figure 5.8a and red in Figure 5.8b. As an example, a closer look to the two datasets in northern London (Figure 5.8b – area 1) is presented in Figure 5.28. However, the additional data are footways, paths or bridleways, which do not comply with the ITN specifications, hence they do not exist in the reference dataset.

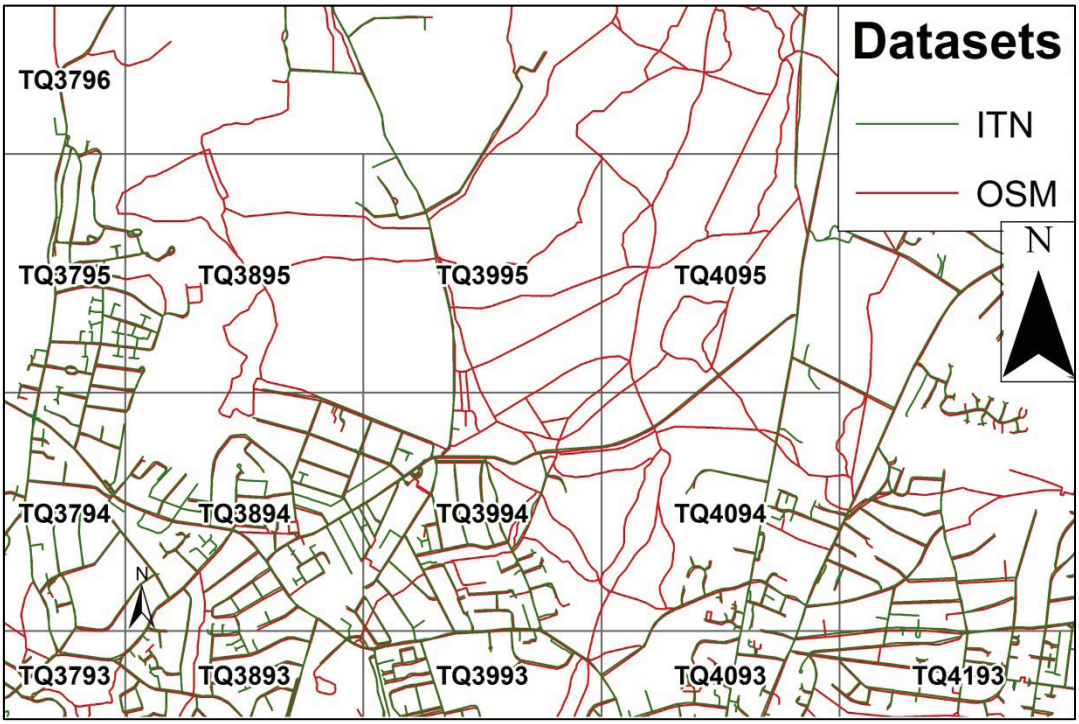

Figure 5.28: *Nothern London: Richer OSM dataset (area 1 of Figure 5.8b)*

Figure 5.8b shows another pattern of richer OSM data for similar reasons, which crosses the London area from east to south. This describes river Thames and its banks, where usually there are no roads close to the river, but there are many pathways, paths, footways. These are only mapped by OSM, hence the reduced OSM matched percentages.

When describing Figure 5.10b, section 5.3 mentioned that OSM secondary names in East London are not present in ITN dataset. A closer manual examination showed that these cases mostly refer to footways and paths, not present in the ITN dataset (OSM non-matched data), and their secondary name values are either numbers (e.g. 40, 61, 82), or text ('Green Chain'). These values are not consistent with the conventional name format (e.g. M25, A14). Thus, they are not cases of VGI

commission, but rather cases of VGI inconsistency, attributed to erroneous tagging of one or more OSM mappers.

Finally, Figure 5.9b shows a linear spatial pattern to the North-East boundary tiles of zero (or close to zero) primary OSM attribute accuracy, meaning that some roads have primary name in the OSM but not in the ITN dataset. A closer look shows that this refers to the M25 London ring road. This conventional name is addressed as secondary one, accurately described by both datasets (Figures 5.10a and b). While there is no primary name (or other official textual description) for M25 (see corresponding blank tiles of Figure 5.9a), OSM users (or maybe just one) appointed a primary name to M25 for these tiles, namely 'London Orbital Motorway'. As a result, this spatial pattern does not prove VGI commission either.

5.5.6. Comparison with results from previous studies

Section 3.4 mentioned Haklay's (2010c) research on OSM in the UK. Particularly in the London area, he visually examined the positional accuracy of five areas (Figure 5.29), sampling 100 roads. Table 5.17 provides his results, next to the average positional accuracy values of this thesis' results for the same areas. The provided methodology gives worse results than Haklay's, with differences up to 7 m. This can be attributed to the fact that the provided algorithm examines all roads instead of a sample. Additionally, results refer to 95% of OSM data, which is a high confidence level. Some features that could otherwise have been considered as outliers (due to their distance) when using a lower confidence level, influence the results by increasing the buffer width. However, the provided results seem more reasonable, taking into consideration the GPS accuracy.

Area	Average Positional Accuracy	Haklay's positional accuracy
Barnet	7.36 m	6.77 m
Highgate	11.19 m	8.33 m
New Cross	13.69 m	6.04 m
South Norwood	10.48 m	3.17 m
Sutton	10.45 m	4.83 m

Table 5.17: *Positional accuracy results compared to Haklay's (2010c)*

Figure 5.29: *OSM and OS's Meridian 2 comparisons across five areas in London (from Haklay, 2010c, p.696)*

Moreover, Haklay (2010c) examined data completeness for the above five areas, also visually. The circles of Figure 5.29 indicate missing data (omission) or digitisation errors in the OSM dataset. Highgate has many minor data omission cases, while Sutton includes large areas of missing data. Additionally to the visual examination, he followed a different approach for data completeness by comparing the datasets' length for each tile of 1 km^2. This divides tiles into three classes: these who have better Meridian coverage (bigger Meridian data length than OSM), these with better OSM coverage and these in the middle. No systematic data matching is performed. Although more detailed results are not provided for a comparison similarly to the one mentioned for positional accuracy, Table 5.18 presents an average ITN matched percentage for each area, as calculated in this study, which shows OSM completeness. The lowest value is in Sutton, which seems in agreement with Haklay (2010c). Further comparison, however, would probably be difficult even if Haklay (2010c) provided more detailed results, firstly because this thesis evaluates all data and, secondly, because since then OSM data coverage is likely to have improved. For example, Highgate appears to be the most complete area, while in Haklay (2010c) was found with many missing (or erroneous) data cases.

Area	Average ITN matched percentages (OSM completeness)
Barnet	96.59%
Highgate	98.28%
New Cross	96.98%
South Norwood	93.16%
Sutton	92.50%

Table 5.18: *OSM completeness provided by this study for the five London areas examined by Haklay (2010c)*

Haklay's (2010c) study on OSM completeness extends to England as a whole. Since the next chapter moves to a national level, also covering his studied area, his results are further discussed in section 6.5.5.

5.6. Summary

The automated method that was presented in Chapter 4 was applied to compare OSM with ITN dataset and provides a systematic approach on VGI spatial quality. The areas selected aim to provide a better understanding of the different quality behaviour of VGI in urban and rural areas and test if the method is suitable for both cases.

Results are in accordance with previous studies, generally showing a better VGI spatial quality (data completeness, positional and attribute accuracy) in urban than in rural areas. Quality results are produced for smaller areas, succeeding in representing VGI heterogeneity efficiently. The approach is systematic, providing more detailed and accurate quality results.

Rural area generally proves to have lower data completeness, attribute and positional accuracy (although secondary name attribute accuracy is unexpectedly higher, as shown in Tables 5.3 and 5.4). The reduced positional accuracy is the reason of slightly bigger errors in data matching, because corresponding objects fail to be considered as such due to their relatively bigger distance. The automated distinction and use of different constraints that was followed helped reducing the problem, however the sporadic nature of such data, occasionally right next to accurate ones, makes it difficult to have a uniform efficiency of the method for both urban and rural areas. Although the evaluation shows that error levels in the automated data matching are almost doubled in rural areas, they still remain below 3.5% for the areas tested here. Successively, further evaluation shows

that error levels of data completeness and attribute accuracy results are quite low (less than 2% and 1% respectively). Positional accuracy results, however, are much more affected by individual objects placed in bigger distances, which along with other factors (data matching errors, different object representation, desired overlap percentage) lead to abnormally low accuracy values, considered as outliers, which may reach 6% of the tiles in rural areas.

The method succeeded in isolating unique features, as well as corresponding features with unique attributes for each dataset, that can be used for conflation purposes. Some manual post-processing was necessary to examine these cases for commissioned data (in other words excessive data that should be present in one dataset but are not).

The urban area of Greater London that was selected for this study includes a complicated and dense road network, which is suitable for checking the efficiency of the method, especially for the data matching part. Official data are specified to be updated and complete, while OSM data are also of a high quality due to the fact that this is where OSM started. This, however, makes it more difficult to check the method efficiency for commissioned data, as few or no real-world objects seem to be left unmapped. The selected rural area, on the other hand, is quite far from OSM home, the network is quite scarce and it is suitable to check for the method efficiency regarding the reduced positional accuracy. However, OSM contribution is also reduced and VGI commission is also difficult to be found. The analysis continues in the next chapter by examining different and larger areas in the UK that include both urban and rural parts (dense or accurate and scarce or less accurate road networks). The efficiency of the method needs also to be tested in such cases. Additionally, larger areas with mixed type of network density and accuracy, where VGI in terms of quality is less likely to be as good as in Greater London, are more likely to show spatial patterns or correlation between quality results.

Chapter 6

Second Case Study: England and Wales

6. Second case study: England and Wales

6.1. Introduction

The previous chapter tested the method in two different but rather uniform areas in terms of density, with results showing that despite its slightly lower efficiency in rural areas, where positional accuracy may be sporadically reduced, the method can effectively apply different constraints for urban and rural areas to achieve data matching with relatively low levels of error. Therefore, the successive quality evaluation, which relies on data matching, gives also quite reliable results.

In this chapter, a far larger area is selected, which includes all of England and Wales. OSM and ITN datasets are again compared as VGI and reference datasets correspondingly. Next sections analyse justification of the studied area, application of the method, results, evaluation and discussion, following a structure similar to the previous chapter.

6.2. Area justification and data preparation

The area selected in this case study aims to test the method for its efficiency in a national scale, with no uniform behavior in terms of road network density, which is also what is likely to occur in a real case scenario. OSM is now systematically tested in areas where there is no previous proof of its quality, far away from where it started. The size of this area and the mixed network density will test if the method works in general. It is also more likely to find spatial differences in quality, which will show if there are any spatial patterns, correlation of quality elements or other findings for VGI.

The same national datasets mentioned in section 5.2 were used, although they were clipped in large areas. The selected regions (Figure 6.1) are something between the first and second national division level, customised to include data that could be processed in a reasonable time. Although the whole studied area could have been processed as one dataset, this was not considered wise in terms of time management, since it would delay the trial-and-error process to improve the developed code. Figure 6.1 and Table 6.1 provides the relevant information. Table 6.1 excludes London region, which is already examined in the previous chapter (see Table 5.1).

Similarly to the first case study, the OSM dataset was reprojected from WGS84 to the BNG, and ITN and OSM datasets were loaded on the PostGIS database using QGIS. The OS 1 km^2 National Grid was used as a tessellation file for each of the above mentioned regions.

Figure 6.1: *Regions examined in the 2nd case study*

Region	Total Tiles (1 km²)	Compared Tiles	Total ITN network length (km)	Total OSM network length (km)
East Anglia	16,869	14,657	40,442.0	42,725.3
Essex	7,959	7,182	25,951.0	31,934.3
Humberside	11,851	9,921	28,049.8	25,649.2
Lancashire	4,461	3,553	18,716.0	17,454.9
Manchester	1,404	1,319	11,949.8	10,875.4
Midlands	10,734	9,998	39,273.0	43,791.7
North	16,394	10,981	34,243.5	31,450.2
Severn	13,018	11,859	38,186.7	37,628.7
South	7,632	7,204	28,938.4	37,295.9
South East	8,044	7,250	25,441.2	28,929.7
South West	17,478	15,391	45,088.4	42,618.5
Wales	22,121	15,443	51,125.5	37,087.6
West	10,833	9,595	28,896.2	33,551.6
Yorkshire	10,732	8,388	27,332.4	26,257.7

Table 6.1: *Studied regions and road network information*

6.3. Method application and results

Using the application described in Appendix A, each region was processed individually. When tiling region borders, datasets were clipped according to the grid and not according to the region border. In this way data extended slightly the region borders, following the tile boundaries, and the border tiles were examined again when the adjacent region was processed (Figure 6.2). However, results were the same because the same data were examined in each case. In this way it was possible to merge the final tile results for each region to produce an overall quality analysis for the whole studied area. Splitting linear datasets according to the region boundaries and then using the normalized grid would result in each border tile having different data size and quality results, depending on the region examined.

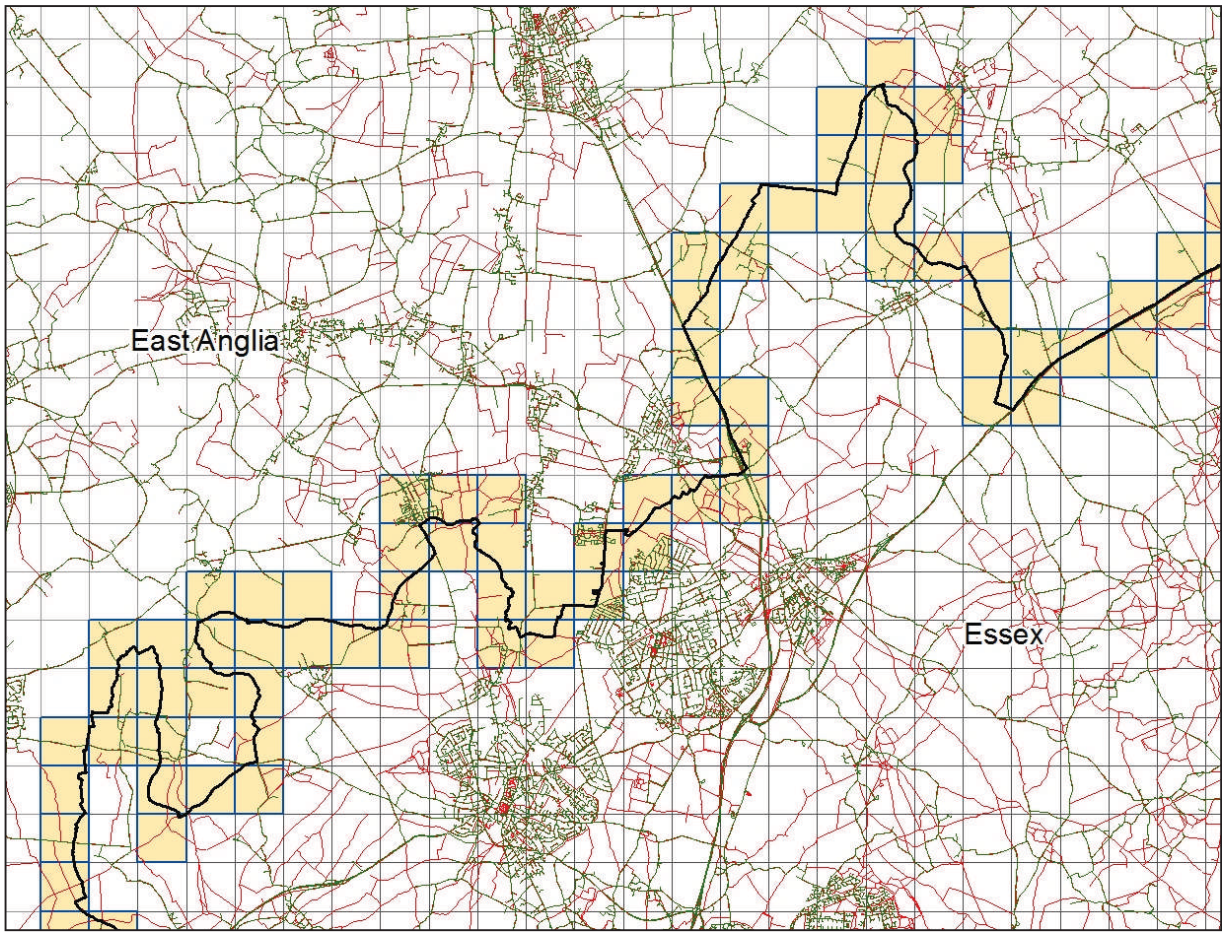

Figure 6.2: *Example of region border tiles and included data, processed individually for each region.*

The method was applied according to the flow diagram of Figure 4.1. Table 6.2 presents the network lengths (total, compared, matched, non-matched) for each dataset and region (see description of Table 5.2 in section 5.3). Again, London region (already presented in Table 5.2) is excluded. (Detailed CSV files that describe each tile individually are produced along with the relevant shapefiles).

Region	Data-set	Total Length (km)	Length Compared (km)	Length Matched (km)	Length non-matched (km)
East Anglia	ITN	40,442.0	39,412.9 (97.46%)	32,927.7 (81.42%)	7,514.3 (18.58%)
	OSM	42,725.3	42,108.4 (98.56%)	32,540.9 (76.16%)	10,184.4 (23.84%)
Essex	ITN	25,951.0	25,821.0 (99.50%)	22,537.1 (86.84%)	3,413.9 (13.16%)
	OSM	31,934.3	31,553.5 (98.81%)	22,087.9 (69.17%)	9,846.4 (30.83%)
Humberside	ITN	28,049.8	27,423.1 (97.77%)	22,696.3 (80.91%)	5,353.5 (19.09%)
	OSM	25,649.2	25,430.9 (99.15%)	22,237.6 (86.70%)	3,411.6 (13.30%)
Lancashire	ITN	18,716.0	18,579.8 (99.27%)	15,415.5 (82.37%)	3,300.5 (17.63%)
	OSM	17,454.9	17,323.5 (99.25%)	14,770.5 (84.62%)	2,684.4 (15.38%)
Manchester	ITN	11,949.8	11,906.3 (99.64%)	9,927.6 (83.08%)	2,022.2 (16.92%)
	OSM	10,875.4	10,835.0 (99.63%)	9,453.3 (86.92%)	1,422.1 (13.08%)
Midlands	ITN	39,273.0	39,074.5 (99.49%)	33,604.9 (85.57%)	5,668.1 (14.43%)
	OSM	43,791.7	43,398.1 (99.10%)	33,065.6 (75.51%)	10,726.1 (24.49%)
North	ITN	34,243.5	32,608.4 (95.22%)	25,344.8 (74.01%)	8,898.7 (25.99%)
	OSM	31,450.2	29,988.5 (95.35%)	24,659.7 (78.41%)	6,790.5 (21.59%)
Severn	ITN	38,186.7	37,289.7 (97.65%)	29,805.6 (78.05%)	8,381.1 (21.95%)
	OSM	37,628.7	37,418.8 (99.44%)	29,301.6 (77.87%)	8,327.2 (22.13%)
South	ITN	28,938.4	28,865.3 (99.75%)	25,063.9 (86.61%)	3,874. 5 (13.39%)
	OSM	37,295.9	36,759.8 (98.56%)	24,656.8 (66.11%)	12,639.1 (33.89%)
South East	ITN	25,441.2	25,291.6 (99.41%)	21,680.5 (85.22%)	3,760.7 (14.78%)
	OSM	28,929.7	28,451.2 (98.35%)	21,233.6 (73.40%)	7,696.1 (26.60%)
South West	ITN	45,088.4	44,565.8 (98.84%)	35,437.5 (78.60%)	9,650.9 (21.40%)
	OSM	42,618.5	41,856.7 (98.21%)	34,717.3 (81.46%)	7,901.2 (18.54%)
Wales	ITN	51,125.5	46,198.3 (90.36%)	32,184.7 (62.95%)	18,940.8 (37.05%)
	OSM	37,087.6	36,210.1 (97.63%)	31,215.6 (84.17%)	5,872.0 (15.83%)
West	ITN	28,896.2	28,615.6 (99.03%)	23,904.6 (82.73%)	4,991.6 (17.27%)
	OSM	33,551.6	32,862.0 (97.94%)	23,483.2 (69.99%)	10,068.4 (30.01%)
Yorkshire	ITN	27,332.4	26,540.5 (97.10%)	20,299.8 (74.27%)	7,032.6 (25.73%)
	OSM	26,257.7	25,254.4 (96.18%)	19,802.9 (75.42%)	6,454.8 (24.58%)

Table 6.2: *Resulting network lengths for study areas*

Figures 6.3 to 6.12 present the quality results. Percentages classification is the same as in the previous chapter, discussed in sections 4.10 and 4.11.3. For positional accuracy a different classification was used, applying buffer intervals that seem to be more suitable for the values distribution of this case study. Generally, higher values (darker green) in ITN figures mean that few or nothing is missing from the OSM dataset and vice versa. Lower values (darker red) in OSM figures mean that some or much data are missing from the ITN dataset (compared to the OSM) and vice versa. More analysis follows in the discussion section 6.5.

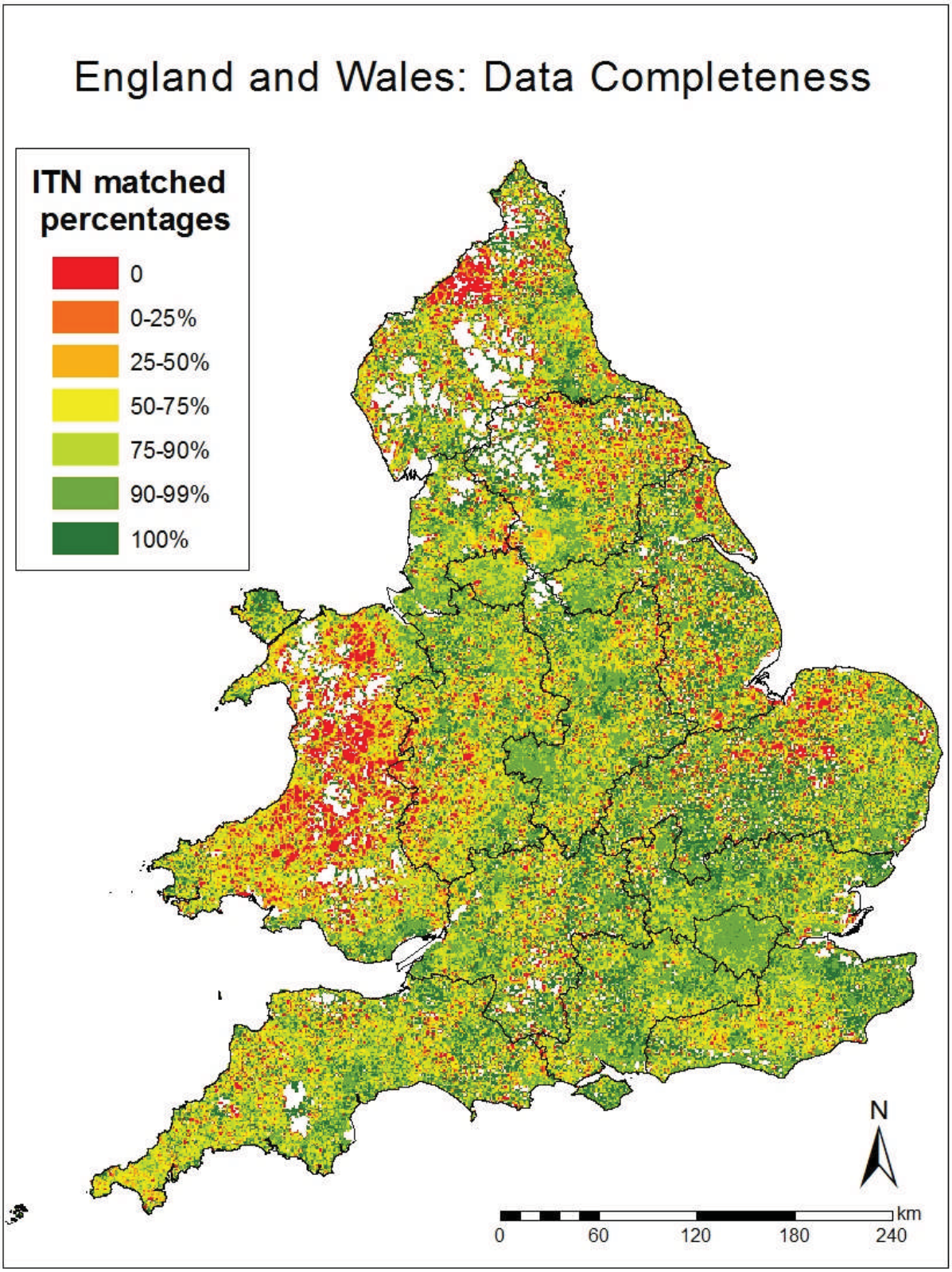

Figure 6.3: *Data Completeness: ITN matched percentages (VGI completeness compared to reference)*

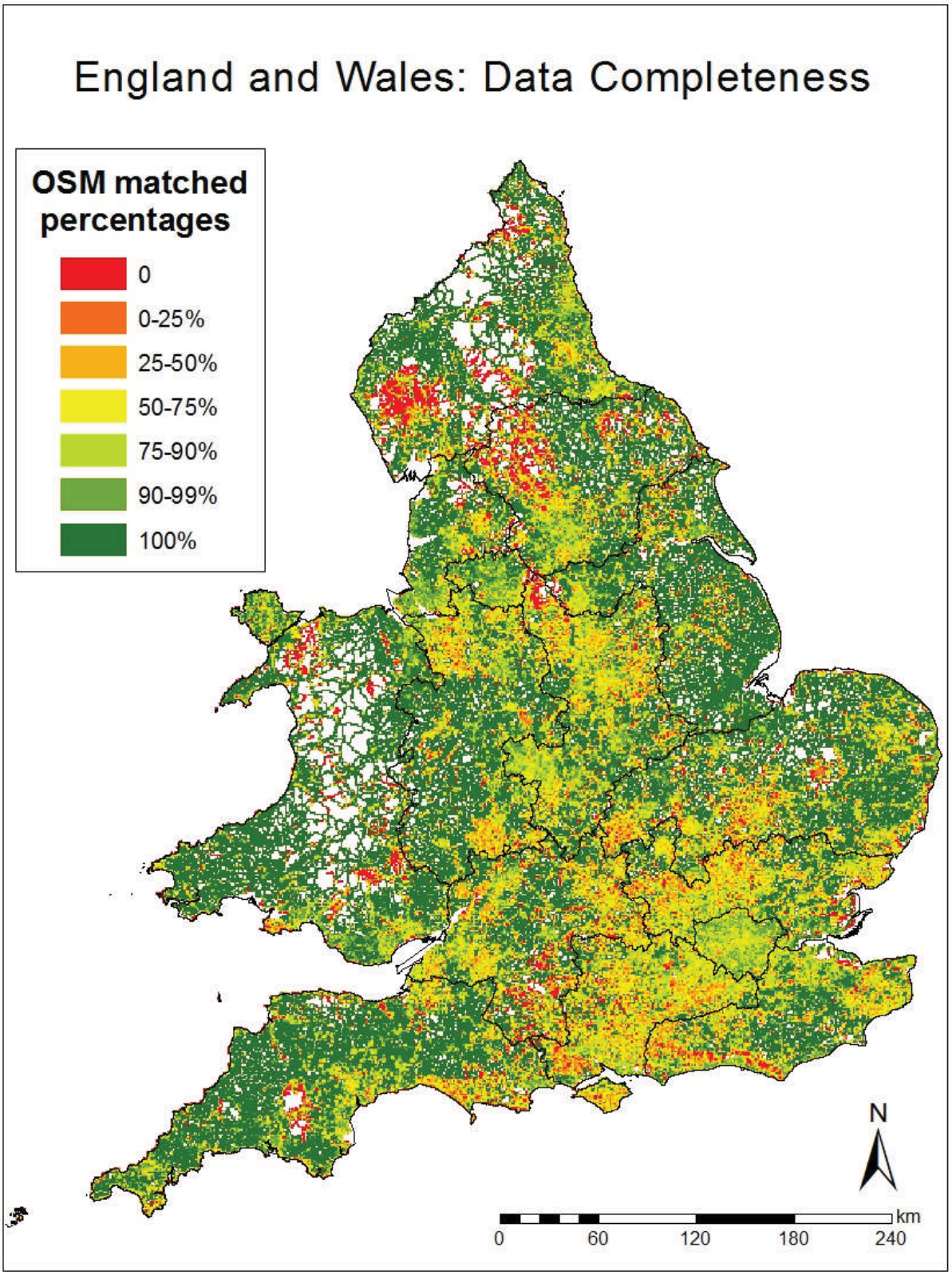

Figure 6.4: *Data Completeness: OSM matched percentages (Red indicate VGI commission)*

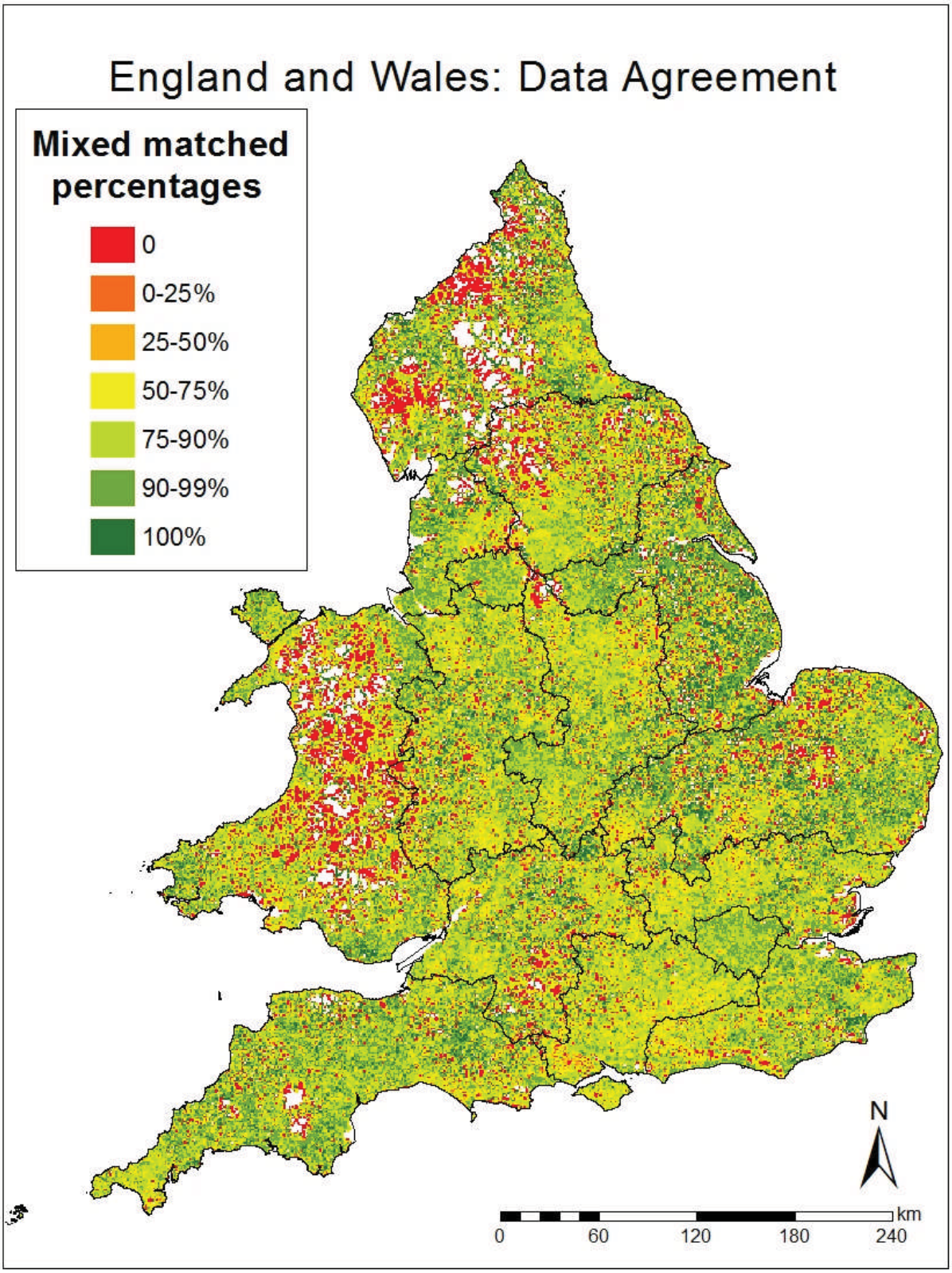

Figure 6.5: *Data Completeness: Level of data agreement (average of ITN and OSM percentages)*

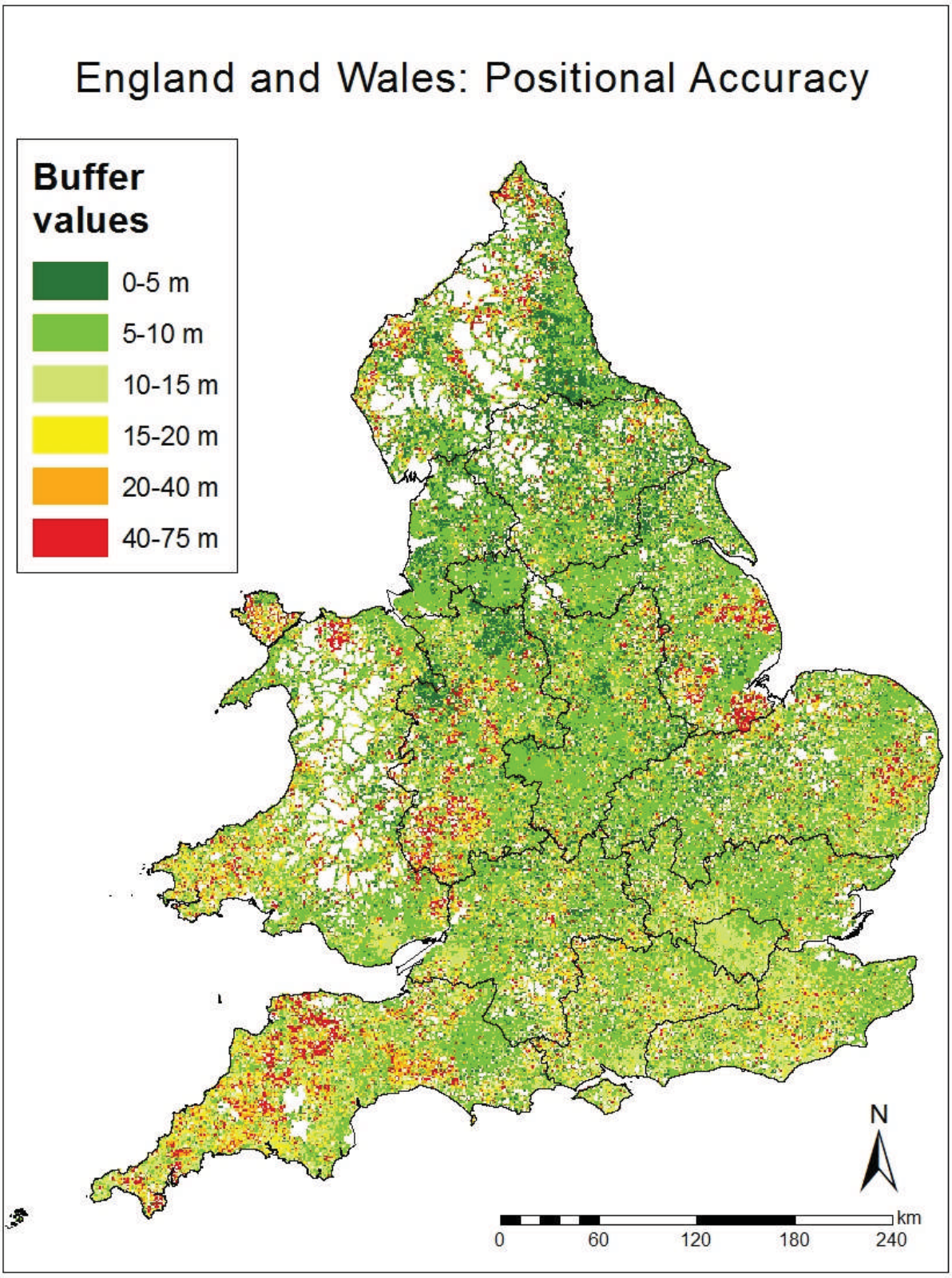

Figure 6.6: *OSM Positional accuracy*

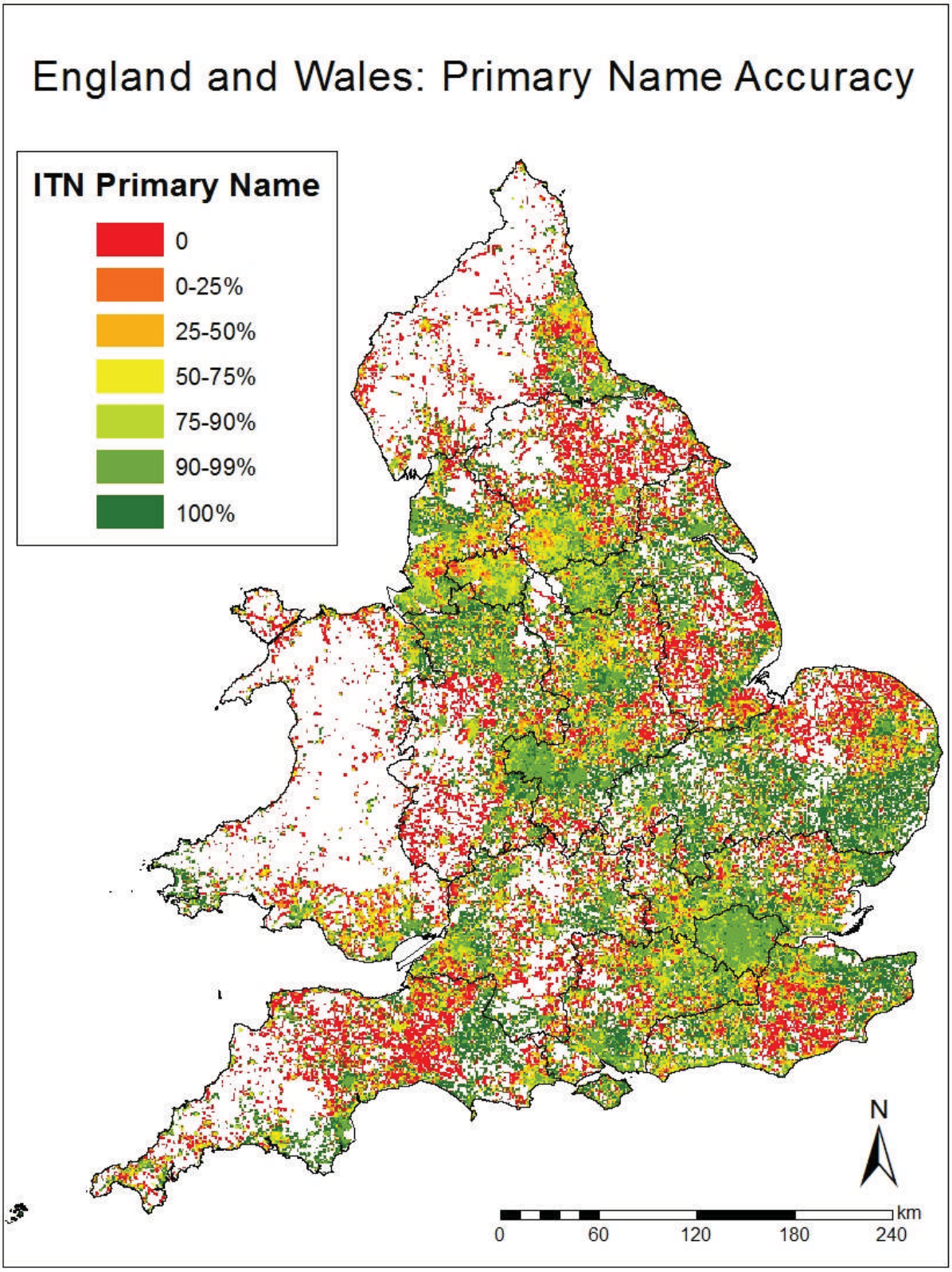

Figure 6.7: *ITN Primary name percentages (OSM primary name accuracy)*

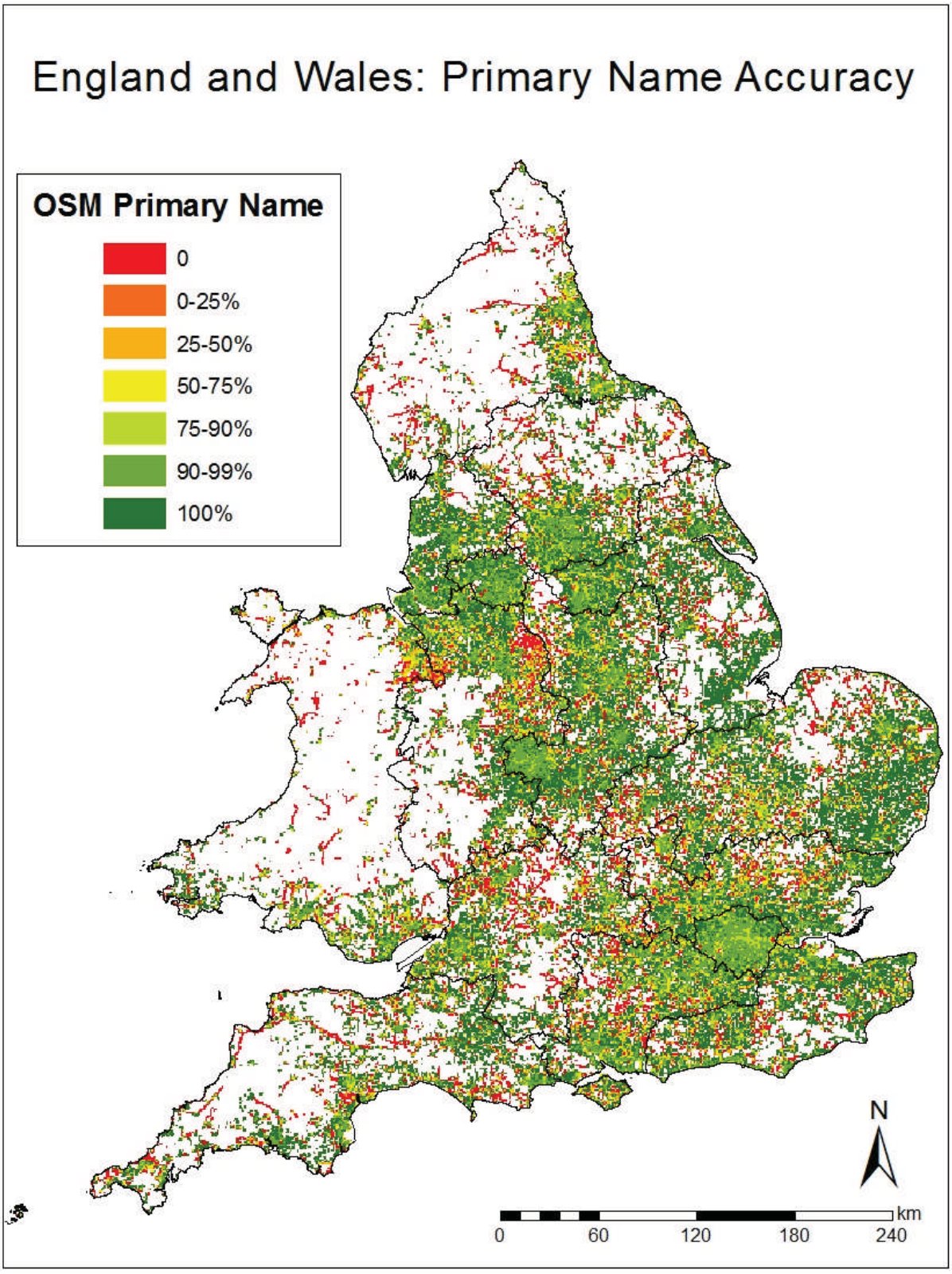

Figure 6.8: *OSM Primary name percentages*

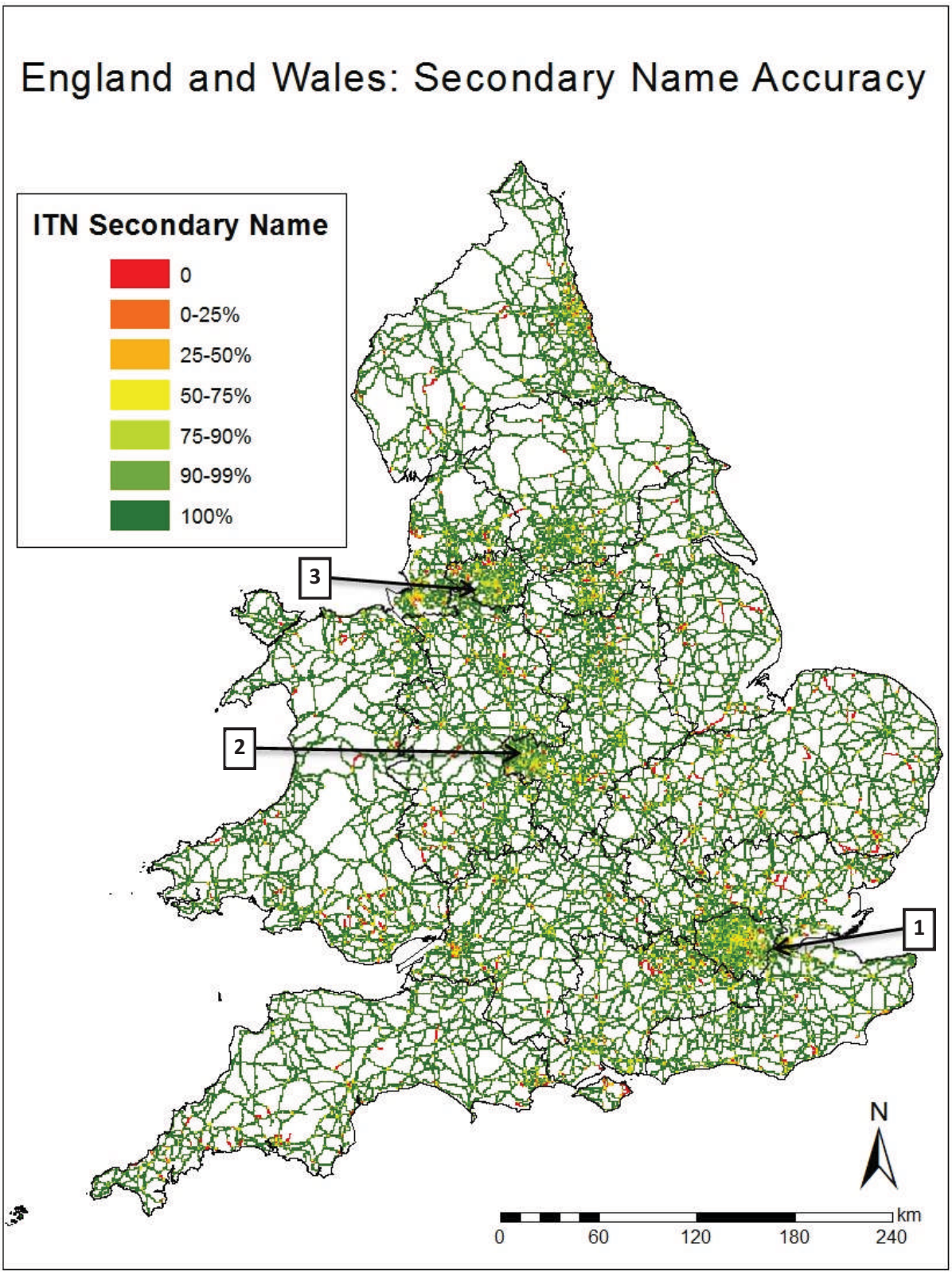

Figure 6.9: *ITN Secondary name percentages*

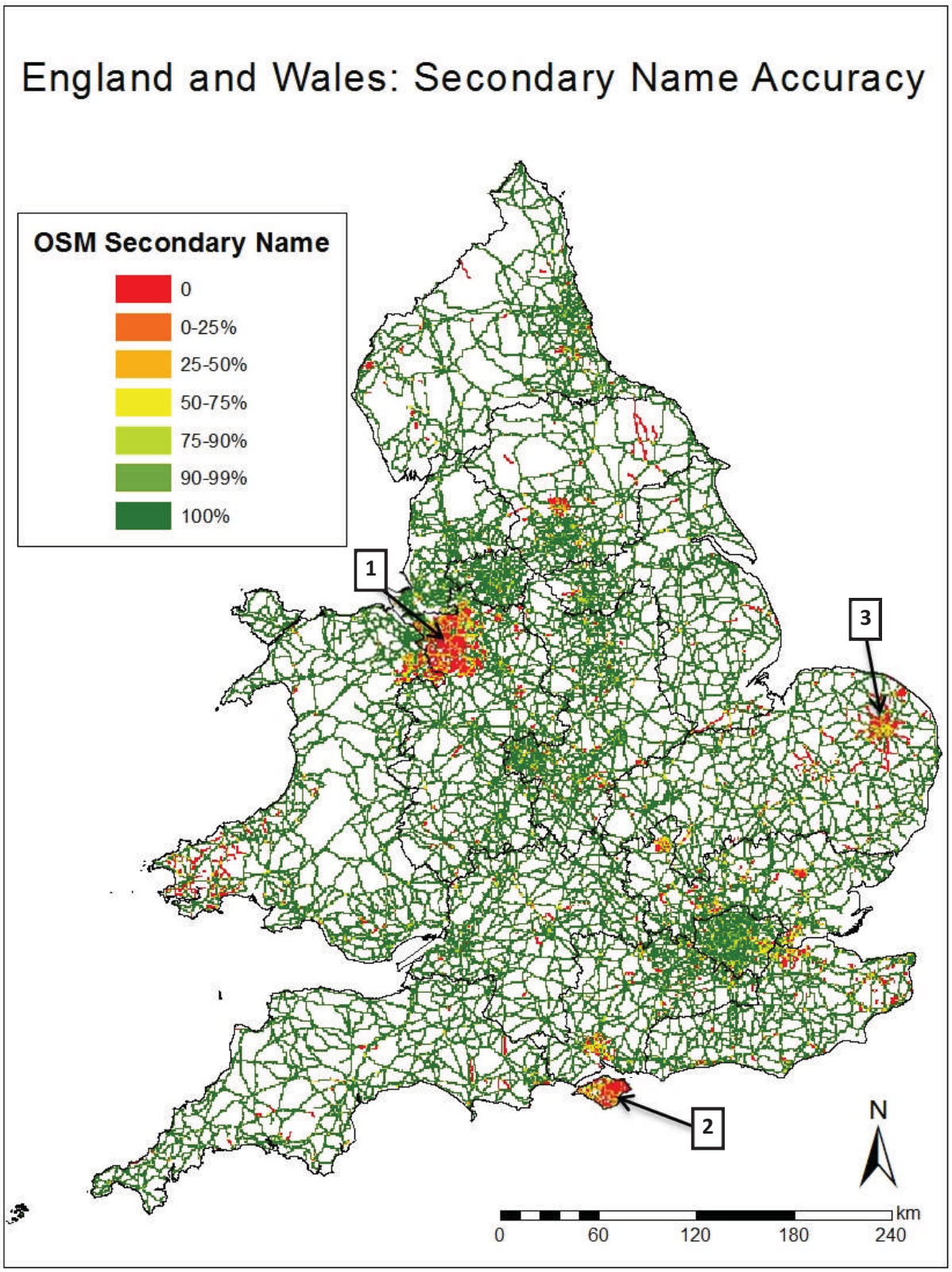

Figure 6.10: *OSM Secondary name percentages*

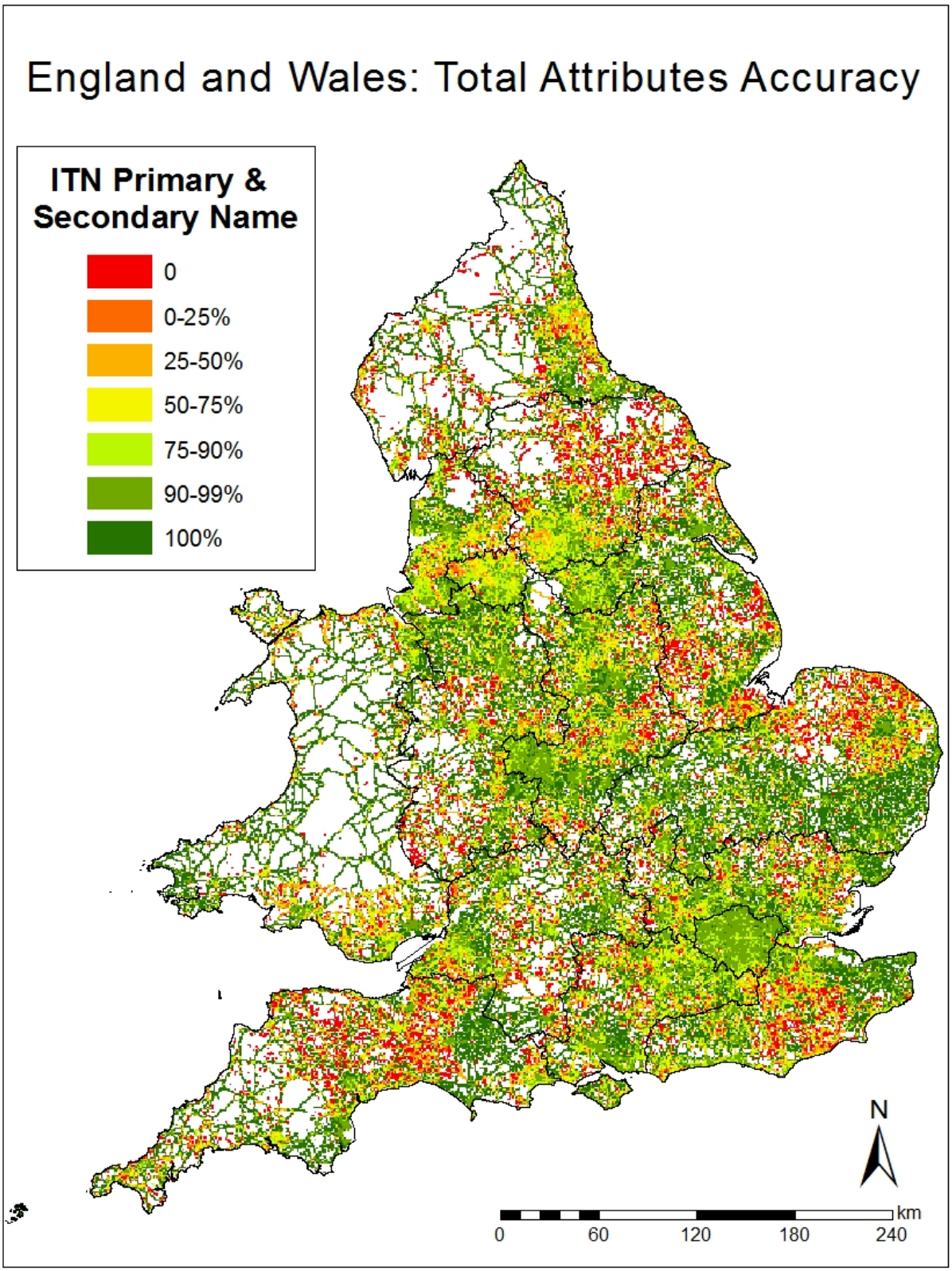

Figure 6.11: *ITN Total attributes accuracy (Primary and secondary name percentages)*

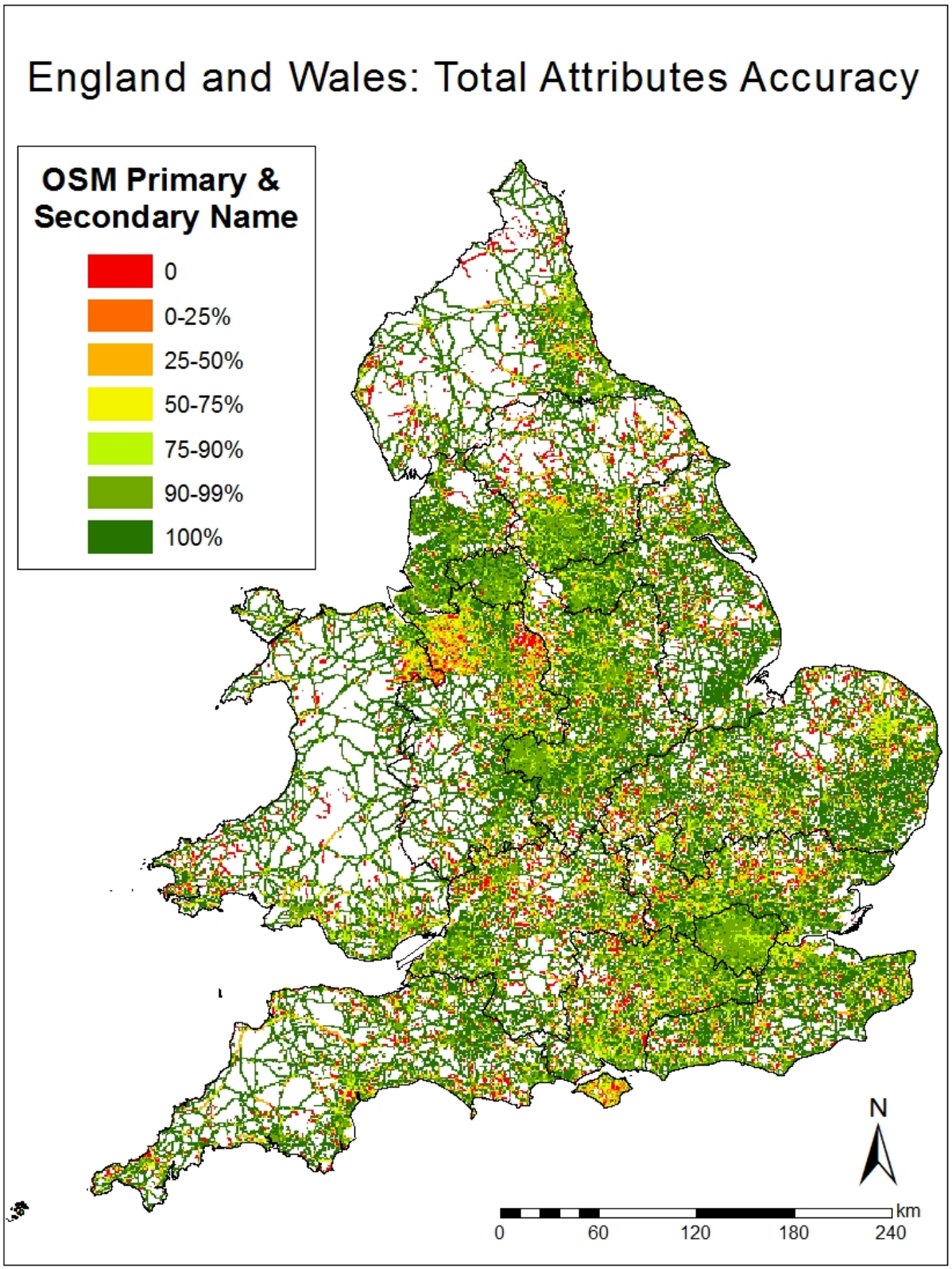

Figure 6.12: *OSM Total attributes accuracy (Primary and secondary name percentages)*

Tables 6.3 to 6.14 provide more statistics regarding the distribution of values per region, similarly to what was discussed in section 5.3 for Tables 5.3 and 5.4 (including the meaning of these values).

East Anglia	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	15,695	75.68	88.50	-1.40	100.00	30.62
	OSM	15,094	79.98	99.00	-1.47	100.00	29.29
Primary name accuracy	ITN	11,770	70.44	96.19	-0.95	100.00	39.47
	OSM	11,147	78.70	100.00	-1.47	100.00	35.62
Secondary name accuracy	ITN	6,076	93.74	100.00	-3.80	100.00	20.63
	OSM	6,341	89.80	100.00	-2.66	100.00	27.93
Total attribute accuracy	ITN	12,569	76.16	97.14	-1.29	100.00	34.45
	OSM	12,213	82.14	100.00	-1.75	100.00	31.14
Positional accuracy (outliers ignored)	OSM	13,642	11.27	8.63	2.92	12.00	9.07

Table 6.3: *Statistics for East Anglia region*

Essex	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	7,340	83.23	92.01	-2.14	99.95	23.43
	OSM	7,453	68.66	74.32	-0.79	98.88	30.15
Primary name accuracy	ITN	5,972	68.71	87.19	-0.93	100.00	37.23
	OSM	5,753	75.81	97.30	-1.28	100.00	35.54
Secondary name accuracy	ITN	3,292	93.20	100.00	-3.63	100.00	20.20
	OSM	3,406	90.19	100.00	-2.79	100.00	26.45
Total attribute accuracy	ITN	6,220	74.02	89.65	-1.26	100.00	33.35
	OSM	6,042	79.75	97.43	-1.59	100.00	31.28
Positional accuracy (outliers ignored)	OSM	6,595	10.86	8.98	3.71	11.50	7.63

Table 6.4: *Statistics for Essex region*

Humberside	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	10,697	74.54	87.25	-1.33	100.00	31.15
	OSM	10,130	88.05	100.00	-2.35	100.00	24.93
Primary name accuracy	ITN	8,262	62.44	84.17	-0.59	100.00	41.61
	OSM	7,080	82.15	100.00	-1.75	100.00	33.81
Secondary name accuracy	ITN	4,231	93.71	100.00	-3.77	100.00	20.44
	OSM	4,218	95.57	100.00	-4.50	100.00	18.62
Total attribute accuracy	ITN	8,710	69.93	88.47	-0.97	100.00	36.76
	OSM	7,913	86.50	100.00	-2.23	100.00	27.82
Positional accuracy (outliers ignored)	OSM	9,412	11.28	8.00	2.62	10.89	10.43

Table 6.5: *Statistics for Humberside region*

Lancashire	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	3,703	77.15	87.53	-1.55	97.02	27.17
	OSM	3,681	81.84	97.42	-1.74	100.00	28.10
Primary name accuracy	ITN	3,184	67.71	82.84	-0.87	98.84	35.53
	OSM	2,973	86.98	100.00	-2.37	100.00	27.98
Secondary name accuracy	ITN	1,906	90.96	100.00	-2.94	100.00	22.07
	OSM	1,873	96.88	100.00	-5.61	100.00	13.77
Total attribute accuracy	ITN	3,234	72.44	84.63	-1.15	98.74	31.57
	OSM	3,110	89.11	100.00	-2.75	100.00	24.77
Positional accuracy (outliers ignored)	OSM	3,352	7.49	6.19	5.60	7.75	6.20

Table 6.6: *Statistics for Lancashire region*

Manchester	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	1,350	78.55	86.06	-1.97	93.01	23.08
	OSM	1,350	81.81	92.23	-1.83	99.77	25.21
Primary name accuracy	ITN	1,289	67.94	79.22	-1.00	92.09	30.62
	OSM	1,246	90.08	98.69	-3.11	100.00	22.24
Secondary name accuracy	ITN	1,026	87.68	100.00	-2.45	100.00	24.67
	OSM	990	97.25	100.00	-6.17	100.00	12.13
Total attribute accuracy	ITN	1,301	72.21	81.21	-1.25	92.05	26.79
	OSM	1,282	91.81	98.75	-3.65	100.00	18.89
Positional accuracy (outliers ignored)	OSM	1,247	7.03	6.00	5.98	7.22	4.76

Table 6.7: *Statistics for Manchester region*

Midlands	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	10,261	79.76	89.82	-1.72	98.86	25.70
	OSM	10,290	74.77	84.15	-1.15	100.00	28.74
Primary name accuracy	ITN	8,532	68.96	89.16	-0.91	100.00	37.79
	OSM	8,009	80.79	99.52	-1.66	100.00	33.35
Secondary name accuracy	ITN	5,059	94.89	100.00	-4.33	100.00	16.21
	OSM	5,106	95.55	100.00	-4.59	100.00	17.56
Total attribute accuracy	ITN	8,917	75.44	91.44	-1.31	100.00	32.33
	OSM	8,557	85.51	99.73	-2.15	100.00	27.05
Positional accuracy (outliers ignored)	OSM	9,363	9.69	7.63	3.66	10.00	7.82

Table 6.8: *Statistics for Midlands region*

North	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	12,525	67.28	80.52	-0.90	98.83	35.21
	OSM	12,151	77.20	100.00	-1.36	100.00	35.07
Primary name accuracy	ITN	4,876	44.70	38.47	0.15	93.58	42.97
	OSM	4,081	63.89	91.75	-0.65	100.00	43.36
Secondary name accuracy	ITN	4,834	95.28	100.00	-4.49	100.00	17.15
	OSM	4,893	95.59	100.00	-4.55	100.00	18.56
Total attribute accuracy	ITN	6,944	69.48	90.76	-0.90	100.00	37.94
	OSM	6,574	81.74	100.00	-1.74	100.00	33.26
Positional accuracy (outliers ignored)	OSM	10,179	11.10	7.05	2.41	11.91	10.90

Table 6.9: *Statistics for North region*

Severn	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	12,661	71.54	82.00	-1.17	95.42	30.08
	OSM	11,994	82.00	100.00	-1.55	100.00	26.71
Primary name accuracy	ITN	7,936	60.80	83.38	-0.52	100.00	42.51
	OSM	7,047	71.75	96.23	-1.04	100.00	39.03
Secondary name accuracy	ITN	5,679	94.83	100.00	-4.28	100.00	17.63
	OSM	6,268	83.98	100.00	-1.86	100.00	33.85
Total attribute accuracy	ITN	9,278	73.80	92.91	-1.19	100.00	35.15
	OSM	8,929	77.99	98.76	-1.40	100.00	33.18
Positional accuracy (outliers ignored)	OSM	11,006	12.77	8.50	2.23	14.25	11.44

Table 6.10: *Statistics for Severn region*

South	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	7,271	83.29	91.21	-2.05	98.91	21.85
	OSM	7,459	64.15	68.90	-0.63	88.92	29.44
Primary name accuracy	ITN	6,034	68.55	86.54	-0.93	100.00	37.02
	OSM	5,861	75.65	96.41	-1.30	100.00	35.71
Secondary name accuracy	ITN	3,512	91.65	100.00	-3.14	100.00	21.79
	OSM	3,707	86.71	100.00	-2.19	100.00	30.38
Total attribute accuracy	ITN	6,331	74.67	88.78	-1.30	100.00	31.90
	OSM	6,229	79.21	96.08	-1.55	100.00	31.06
Positional accuracy (outliers ignored)	OSM	6,403	12.73	10.50	3.30	13.63	8.41

Table 6.11: *Statistics for South region*

South East	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	7,457	80.00	88.95	-1.78	98.42	24.95
	OSM	7,516	74.00	85.39	-1.11	100.00	30.32
Primary name accuracy	ITN	6,320	62.12	81.13	-0.57	100.00	40.70
	OSM	5,545	80.06	100.00	-1.59	100.00	33.94
Secondary name accuracy	ITN	3,243	95.49	100.00	-4.55	100.00	15.17
	OSM	3,436	90.49	100.00	-2.84	100.00	26.33
Total attribute accuracy	ITN	6,496	69.38	85.64	-0.93	100.00	35.77
	OSM	6,046	83.62	100.00	-1.93	100.00	28.90
Positional accuracy (outliers ignored)	OSM	6,649	12.59	10.52	3.47	13.56	7.95

Table 6.12: *Statistics for South East region*

South West	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	16,008	75.05	83.39	-1.37	96.15	26.89
	OSM	16,026	82.57	100.00	-1.73	100.00	29.09
Primary name accuracy	ITN	9,147	50.21	56.77	-0.05	100.00	44.34
	OSM	7,310	70.93	99.40	-0.97	100.00	40.83
Secondary name accuracy	ITN	5,740	95.26	100.00	-4.43	100.00	17.95
	OSM	5,743	96.44	100.00	-5.13	100.00	16.76
Total attribute accuracy	ITN	10,779	64.90	84.53	-0.68	100.00	40.03
	OSM	9,543	81.74	100.00	-1.71	100.00	32.54
Positional accuracy (outliers ignored)	OSM	14,204	16.12	11.88	1.85	19.94	10.95

Table 6.13: *Statistics for South West region*

Wales	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	18,770	55.43	63.30	-0.40	87.28	35.80
	OSM	16,224	84.87	100.00	-2.02	100.00	29.52
Primary name accuracy	ITN	5,530	46.50	45.79	0.08	94.62	42.32
	OSM	4,844	61.62	85.20	-0.54	100.00	42.85
Secondary name accuracy	ITN	7,316	95.33	100.00	-4.50	100.00	17.78
	OSM	7,630	91.22	100.00	-2.95	100.00	26.19
Total attribute accuracy	ITN	9,218	76.69	99.25	-1.28	100.00	33.42
	OSM	9,308	81.45	100.00	-1.68	100.00	32.88
Positional accuracy (outliers ignored)	OSM	14,291	13.12	9.61	2.40	14.75	10.29

Table 6.14: *Statistics for Wales*

West	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	9,928	78.65	88.90	-1.64	100.00	27.29
	OSM	10,050	71.41	82.48	-0.93	100.00	32.07
Primary name accuracy	ITN	6,525	60.68	80.97	-0.50	100.00	41.84
	OSM	6,228	68.31	94.66	-0.85	100.00	40.86
Secondary name accuracy	ITN	4,423	95.72	100.00	-4.69	100.00	15.92
	OSM	4,439	97.24	100.00	-5.97	100.00	14.32
Total attribute accuracy	ITN	7,491	72.64	91.86	-1.10	100.00	35.56
	OSM	7,320	78.37	99.25	-1.46	100.00	33.70
Positional accuracy (outliers ignored)	OSM	8,738	11.49	9.00	3.16	12.81	8.63

Table 6.15: *Statistics for West region*

Yorkshire	Dataset	Tiles evaluated	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	ITN	9,173	68.87	79.53	-0.98	96.99	32.48
	OSM	9,175	74.03	94.86	-1.14	100.00	34.79
Primary name accuracy	ITN	6,518	47.03	48.07	0.04	92.75	42.08
	OSM	5,150	72.71	98.09	-1.09	100.00	39.36
Secondary name accuracy	ITN	3,287	96.03	100.00	-5.08	100.00	14.81
	OSM	3,476	91.31	100.00	-2.99	100.00	26.18
Total attribute accuracy	ITN	6,948	57.68	68.79	-0.41	97.71	39.53
	OSM	6,005	78.99	99.66	-1.52	100.00	33.91
Positional accuracy (outliers ignored)	OSM	7,705	9.44	6.66	3.35	9.75	9.17

Table 6.16: *Statistics for Yorkshire region*

6.4. Evaluation

6.4.1. Object matching efficiency

Section 5.4.2 suggested that a manual evaluation of around 5% of the tiles is enough to assess the errors of the automation procedure. When this case study has 156,729 tiles, however, this means 7,836 tiles, which is still a forbidding amount of manual work. In such cases of large populations, statistical theory suggests smaller sample sizes. What needs to be considered is the population distribution. Quality results may not follow a normal distribution, as data completeness and accuracy depend on the number of users, their dedication and the importance of the area. Data matching errors, on the other hand, do not depend on the above factors, as the automation guarantees a uniform behavior. Hence, they are more likely to follow a normal distribution or something similar,

provided that the tiles that will be evaluated are appropriately distributed to include areas of high as well as low completeness and accuracy. Student's t-distribution, for example, suggests that an appropriate sample size should be above 30: Student's t-test results do not change significantly when sample size > 30. In this context, it was decided to evaluate 120 tiles in total, divided in 8 tiles per region. These tiles are semi-randomly selected for each region and the following rules are followed: They should be well distributed and three out of eight per region should refer to urban areas – dense network. This is because in previous chapter it was found that rural areas have a reduced quality, so by evaluating only urban tiles evaluation results might be more optimistic. Since evaluation relies on network length, a three to five relation between tiles of urban and rural areas respectively gives a more balanced result. Feature matching evaluation follows the method described in section 5.4.2. Figure 6.13 presents the tiles that were manually evaluated. The tiles evaluated in the previous chapter for the London region (Figure 5.18) are not included in the above selection (different tiles are selected).

Figure 6.13: *Tiles manually evaluated*

Region	Data-set	Length (m) evaluated	Missing length (m)	Surplus length (m)	Total matching error (m)
East Anglia	ITN	51,457.0	0.0 (0%)	270.8 (0.53%)	270.8 (0.53%)
	OSM	54,231.3	18.9 (0.03%)	228.1 (0.42%)	247.0 (0.46%)
Essex	ITN	59,784.7	0.0 (0%)	1,622.3 (2.71%)	1,622.3 (2.71%)
	OSM	89,225.5	336.3 (0.38%)	1,262.4 (1.41%)	1,598.7 (1.79%)
Humberside	ITN	58,367.8	80.5 (0.14%)	1,198.1 (2.05%)	1,278.6 (2.19%)
	OSM	55,952.9	89.4 (0.16%)	138.5 (0.25%)	227.9 (0.41%)
Lancashire	ITN	58,202.8	56.0 (0.1%)	911.7 (1.57%)	967.7 (1.66%)
	OSM	54,881.9	287.8 (0.52%)	101.3 (0.18%)	389.1 (0.71%)
London	ITN	70,014.6	640.7 (0.92%)	2,009.1 (2.87%)	2,649.8 (3.78%)
	OSM	80,465.5	765.9 (0.95%)	20.2 (0.03%)	786.1 (0.98%)
Manchester	ITN	72,676.7	0.0 (0%)	3,895.2 (5.36%)	3,895.2 (5.36%)
	OSM	61,419.3	258.4 (0.42%)	117.5 (0.19%)	375.9 (0.61%)
Midlands	ITN	56,138.8	110.8 (0.2%)	939.1 (1.67%)	1,049.9 (1.87%)
	OSM	56,173.5	0.0 (0%)	35.9 (0.06%)	35.9 (0.06%)
North	ITN	49,629.0	696.4 (1.4%)	912.3 (1.84%)	1,608.7 (3.24%)
	OSM	47,868.2	0.0 (0%)	0.0 (0%)	0.0 (0%)
Severn	ITN	51,637.0	137.4 (0.27%)	1,030.6 (2%)	1,168.0 (2.26%)
	OSM	58,497.3	131.6 (0.22%)	446.8 (0.76%)	578.4 (0.99%)
South	ITN	45,733.7	0.0 (0%)	631. 7 (1.38%)	631.7 (1.38%)
	OSM	55,203.8	139.5 (0.25%)	411.4 (0.75%)	550.9 (1%)
South East	ITN	50,895.7	0.0 (0%)	466.6 (0.92%)	466.6 (0.92%)
	OSM	53,756.3	131.9 (0.25%)	973.4 (1.81%)	1,105.3 (2.06%)
South West	ITN	19,854.0	0.0 (0%)	226.2 (1.14%)	226.2 (1.14%)
	OSM	19,988.8	0.0 (0%)	36.2 (0.18%)	36.2 (0.18%)
Wales	ITN	32,075.5	13.6 (0.04%)	1,313.5 (4.09%)	1,327.1 (4.14%)
	OSM	32,542.2	161.3 (0.5%)	32.2 (0.1%)	193.5 (0.59%)
West	ITN	53,340.8	0 (0%)	768.5 (1.44%)	768.5 (1.44%)
	OSM	62,851.0	524.8 (0.84%)	674.4 (1.07%)	1,199.2 (1.91%)
Yorkshire	ITN	54,551.8	0.0 (0%)	1,345.8 (2.47%)	1,345.8 (2.47%)
	OSM	53,581.0	0.0 (0%)	88.7 (0.17%)	88.7 (0.17%)
Total	**ITN**	**784,359.7**	**1,735.5 (0.22%)**	**17,541.5 (2.24%)**	**19,276.9 (2.46%)**
	OSM	**836,638.6**	**2,845.9 (0.34%)**	**4,566.9 (0.55%)**	**7,412.7 (0.89%)**

Table 6.17: *Data matching errors (per region – dataset and total)*

Table 6.17 presents the manual evaluation results (for each region as well as for the whole area studied). Figure 6.14 shows the total error percentage in respect to the number of tiles examined, which seems to stabilize above 80 tiles for both datasets.

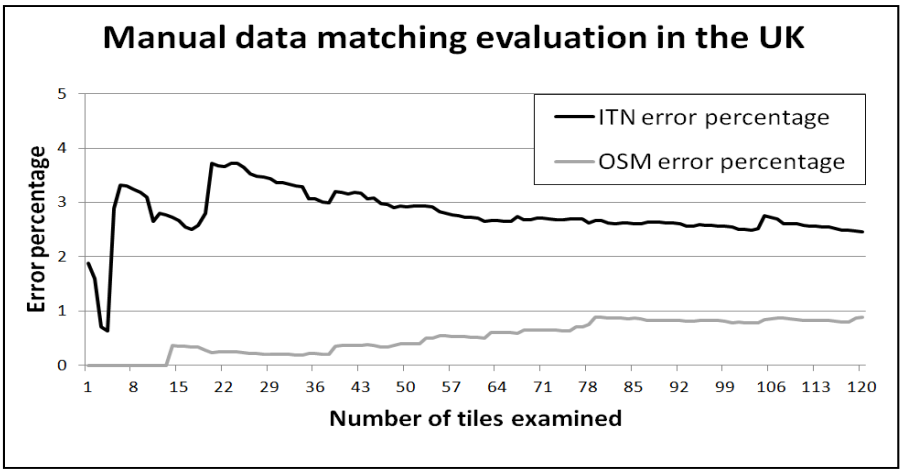

Figure 6.14: *Data matching error compared to number of tiles evaluated*

Compared to the manual evaluation results of the previous case study that separated urban and rural areas (see Table 5.6), the ITN data matching errors agree with the assumption (made in section 5.5.1) that in a general case with mixed network density error levels are expected to be between the ones found in the previous chapter. The OSM error levels found in this case study, however, are even lower than expected.

6.4.2. Attribute accuracy efficiency

Section 4.11 described the method to measure attribute accuracy. Sections 4.14 and 5.4.3 described how to manually evaluate attribute accuracy results, distinguishing three types of error: type 1 for objects mistakenly considered as accurate due to erroneous data matching, type 2 for failure in text similarity, resulting in accepting as accurate two different names, and type 3 for failure in text similarity in the opposite way. This case study follows the same procedure for 84 out of the 120 tiles of Figure 6.13 (Figure 6.15). Error levels (Table 6.18) are slightly higher than in previous chapter, however they still remain quite low (less than 2.5%).

Figure 6.15: *Tiles manually evaluated for attribute accuracy*

England and Wales	ITN matched dataset		OSM matched dataset	
	Length (m)	Pct (%)	Length (m)	Pct (%)
Total attributes	139,676.7	100.00	120,774.2	100.00
Primary name	107,152.2	76.71	86,960.8	72.00
Secondary name	32,524.6	23.29	33,813.5	28.00
Error type 1	2,039.9	1.46	2,325.4	1.93
Error type 2	112.1	0.08	0.000	0.00
Error type 3	90.6	0.06	372.5	0.31
Total errors	2,242.6	**1.61**	2,697.9	**2.23**

Table 6.18: *Attribute accuracy errors*

6.4.3. Positional accuracy efficiency

Table 6.19 provides information on the number of tiles considered as outliers in terms of positional accuracy (explained in sections 4.12.4 and 5.4.4) for each region. They range from 2.8% to 9.2% of the total tiles examined in each region. Figure 6.16 shows that outliers form some sort of clusters that need to be examined. It may not be by chance that the first four highest percentages of outliers refer to adjacent regions (Table 6.23).

Region	Tiles Buffered	Outliers
East Anglia	14,345	703 (4.90 %)
Essex	7,036	441 (6.27 %)
Humberside	9,743	331 (3.40 %)
Lancashire	3,491	139 (3.98 %)
London	1,681	47 (2.80 %)
Manchester	1,300	53 (4.08 %)
Midlands	9,786	423 (4.32 %)
North	10,658	479 (4.49 %)
Severn	11,638	631 (5.42 %)
South	7,052	649 (9.20 %)
South East	7,102	453 (6.38 %)
South West	15,124	920 (6.08 %)
Wales	15,029	737 (4.90 %)
West	9,312	574 (6.16 %)
Yorkshire	8,147	442 (5.43 %)
Total	**112,820**	**6,801 (5.49%)**

Table 6.19: *Outliers found in each region*

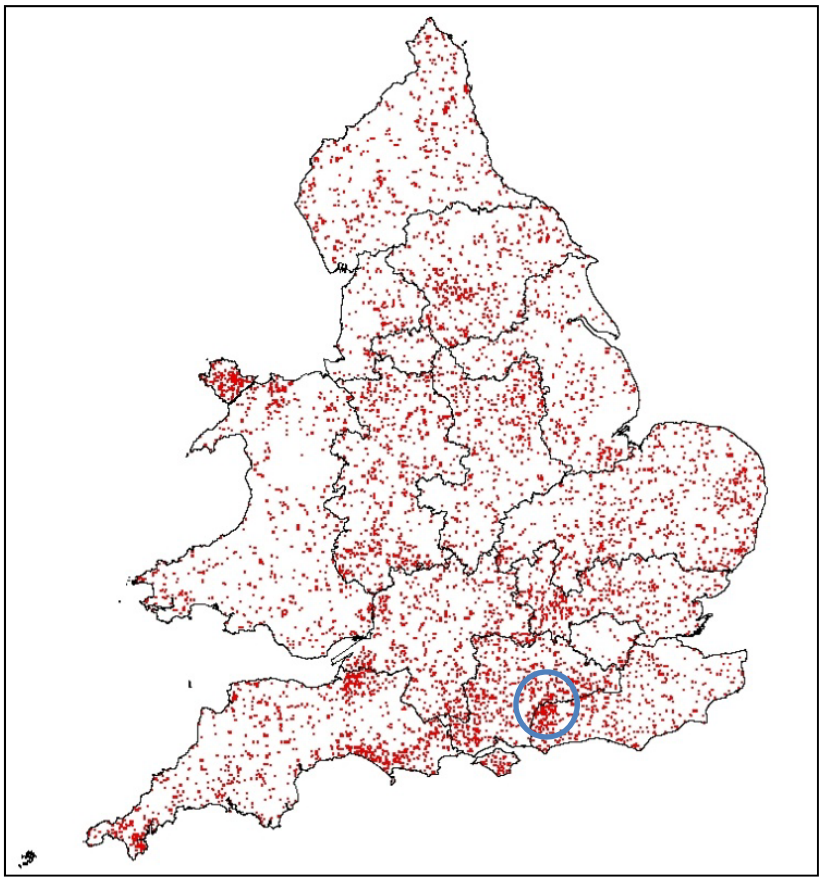

Figure 6.16: *Tiles considered as outliers for positional accuracy and evaluated area*

An area West of the South-East region (also extending to the South region, see blue circle in Figure 6.16) with concentrated outliers was manually examined, following the method described in section 5.4.4, to comprehend what usually leads to such large buffer sizes. Table 6.20 presents the results.

	Number	Percentage
Total outliers	6,801	4.65% of total tiles
Outliers examined:	350	5.15% of total outliers
Data matching errors	112	32.00%
Distance errors	45	12.86%
Different representation	136	38.86%
Topology errors	57	16.29%

Table 6.20: *Outliers manually examined*

All the examined outliers were found in areas with a rather low network density (rural areas), classified either as rural or as urban due to the low threshold used for the distinction (see section 4.7). Additionally to what is described in section 5.4.4, a fourth reason was found that leads to outliers. 'Topology errors' of Table 6.20 are actually cases of 'Different representation' of features (explained in section 5.4.4). However, in this case the different representation that leads to an extended VGI length or partially different shape, refers to objects that are not divided at road junctions when intersected. While the simple 'Different representation' of Figure 5.23b leads to increased buffers that cannot be predicted or avoided, the one attributed to 'Topology errors' could be minimised by correcting the topology. However, despite their significant contribution to the total number of outliers in the above sample, the most important factor remains the simple 'Different representation'. Data matching errors, as a reason for outliers, comes next, which means that the 'Outlier' indication could efficiently serve as a quality measure tool to find and correct data matching errors. 'Distance errors', which refer to tiles with low positional accuracy due to an increased distance of corresponding objects, hence should not be considered as outliers, come last as a factor for outliers. This, similarly to the previous case study (Table 5.9), proves the efficiency of the thresholds used to define the outliers in section 4.12.4.

6.4.4. VGI commission indication

Section 5.4.5 explained how the indication for new objects in the OSM dataset is manually evaluated. As already noted there, the evaluation refers to the classification of a non-matched OSM feature as new regardless of its road type, in other words a footpath that does not (and should not)

exist in the ITN dataset will be marked as a new object in the OSM dataset. The results for the tiles of Figure 6.13 (Table 6.21) show low levels of erroneous 'new object' indication.

Evaluation of VGI Commission indication	OSM dataset: England and Wales	
	Number	Length
Evaluated non-matched objects	1,235	149,209.3 (100%)
Indicated non-matched objects as new	377	436,17.6 (29.23%)
Data matching errors (instead of new objects)	14	2054.1 **(1.38%)**

Table 6.21: *Evaluation of VGI commission indication*

6.5. Discussion

6.5.1. Data matching errors and quality results

Section 4.14 described how errors in data matching affect the measurement of quality elements, using simple equations. Section 5.5.1 applied these equations on the errors found during the manual evaluation of data matching. Similarly in this case study, by using the results from Tables 6.17, 6.18, 6.20 and equations 11, 12 and 14 of section 4.14, Table 6.22 gives an estimation of the quality results' errors due to erroneous data matching. Positive values regarding OSM completeness and attribute accuracy mean that on average results are more optimistic (better) than they should be.

VGI spatial quality element	Estimated error range
OSM Completeness (based on data matching)	On average +2.2% (-0.44 % to +5.36 % per region)
Attribute accuracy	On average +1.48 %
Positional accuracy	Up to 1.49% of outliers

Table 6.22: *Estimation of errors in quality results for the provided method*

6.5.2. Correlation of quality results: spatial patterns

Combining the results of Figures 6.3 to 6.12 and Tables 6.3 to 6.16, some interesting conclusions can be drawn. Table 6.23 presents the average quality values for all regions, highlighting in green the maximum and in light red the minimum average values.

Regions	ITN percentages				OSM percentages				OSM Positional accuracy (m)
	Data match	1mary name	2ndary name	Total attributes	Data match	1mary name	2ndary name	Total attributes	
East Anglia	75.68	70.44	93.74	76.16	79.98	78.70	89.80	82.14	11.27
Essex	83.23	68.71	93.20	74.02	68.66	75.81	90.19	79.75	10.86
Humberside	74.54	62.44	93.71	69.93	88.05	82.15	95.57	86.50	11.28
Lancashire	77.15	67.71	90.96	72.44	81.84	86.98	96.88	89.11	7.49
London	91.76	90.99	90.31	91.17	78.77	91.56	93.32	92.31	11.38
Manchester	78.55	67.94	87.68	72.21	81.81	90.08	97.25	91.81	7.03
Midlands	79.76	68.96	94.89	75.44	74.77	80.79	95.55	85.51	9.69
North	67.28	44.70	95.28	69.48	77.20	63.89	95.59	81.74	11.10
Severn	71.54	60.80	94.83	73.80	82.00	71.75	83.98	77.99	12.77
South	83.29	68.55	91.65	74.67	64.15	75.65	86.71	79.21	12.73
South East	80.00	62.12	95.49	69.38	74.00	80.06	90.49	83.62	12.59
South West	75.05	50.21	95.26	64.9	82.57	70.93	96.44	81.74	16.12
Wales	55.43	46.50	95.33	76.69	84.87	61.62	91.22	81.45	13.12
West	78.65	60.68	95.72	72.64	71.41	68.31	97.24	78.37	11.49
Yorkshire	68.87	47.03	96.03	57.68	74.03	72.71	91.31	78.99	9.44

Table 6.23: *Average quality values per region, highlighting highest (green) and lowest (red) scores*

Greater London, the birth place of OSM, has the maximum average ITN matching percentage, primary name and total attribute percentage, which means that it is the region where OSM is most complete in data and most accurate in attributes, compared to the reference dataset. Interestingly, however, positional accuracy is rather disappointing compared to other regions (11.38m, Table 5.3 – Figure 6.6, which brings London to the ninth place out of 15 regions examined). Manchester, which is the most accurate in terms of position (7.03m), has the seventh position in OSM data completeness (ITN matched percentage), the tenth in total attribute accuracy and the worse in secondary road name accuracy. OSM in Yorkshire has the highest quality in secondary name, but the second worst quality in primary name, which leads the region to the last place for total attribute accuracy. These examples show that correlations between quality elements, if they exist, are not obvious, and appropriate statistical tools may need to be applied to assess them.

However, some spatial patterns can be distinguished in Figures 6.3 to 6.12. For example, although Wales is the region with the worst score regarding OSM completeness (Figure 6.3 and Table 6.23), it includes smaller areas where OSM is richer in data. Generally OSM is richer in South England and around Manchester (lower matched percentages in Figure 6.4). Apart from data completeness, better total attribute accuracy (primary and secondary name) seems to apply to the same areas. Northern mountainous areas (such as the area of Lake District, Figure 6.21) are also far better

covered by OSM, but this is due to the types of roads, which are not meant for vehicles and therefore not mapped by the ITN dataset (paths, bridleways, footways, tracks). Other visible spatial patterns do not follow region borders but usually occur in smaller areas, which can lead to questioning the region by region evaluation, since average quality statistics are not really representative. Since such patterns are unknown a-priori, this problem is likely to occur when using administrative boundaries. The fact that the method is automated, however, enables the quality evaluation repetition, using different areas to produce more representative statistics per area.

Because of the unexpected lower positional accuracy in London, the area was further examined. Section 5.5.3 described the use of satellite imagery along with the output datasets. Figure 5.25 that was used in previous chapter for London area, as well as Figure 6.17 regarding ITN and OSM matched data, show that there is a rather systematic dislocation in OSM dataset in the area of London, which is around 8 m, compared to the ITN and satellite imagery. This leads to increased buffer values and lower positional accuracy. Figure 6.17 uses Yahoo! instead of Google imagery, since Yahoo provides the satellite imagery for OSM, however the visual result is the same regardless of the background base map provider, which rather rules out a case of error in Yahoo satellite imagery positioning.

Figure 6.17: *OSM (red) and ITN (yellow) matched data in London area: OSM dislocation to the south, (Satellite imagery source: Bing Maps provided by ESRI's ArcMap)*

Secondary name accuracy is high for all regions, which means that the rural network is well-mapped and tagged. A few exceptions in the urban areas of London, Birmingham, Manchester (Figure 6.9 – areas 1, 2, 3 respectively), show that the secondary official name is not always captured or known to the OSM mappers. On the other hand, the large areas of low (red) values when comparing OSM with ITN (Figure 6.10 – areas 1,2) theoretically show OSM commission (attribute information non-existing in ITN), however it is highly unlikely to happen to such an extent. A closer look in the northern Severn region (Figure 6.18 / area 1 of Figure 6.10) shows that OSM uses unusual secondary names (displayed in red text), which can be attributed to a specific user that does not follow OSM suggestions on tagging. As a result, such clusters could be an indication of OSM structural errors.

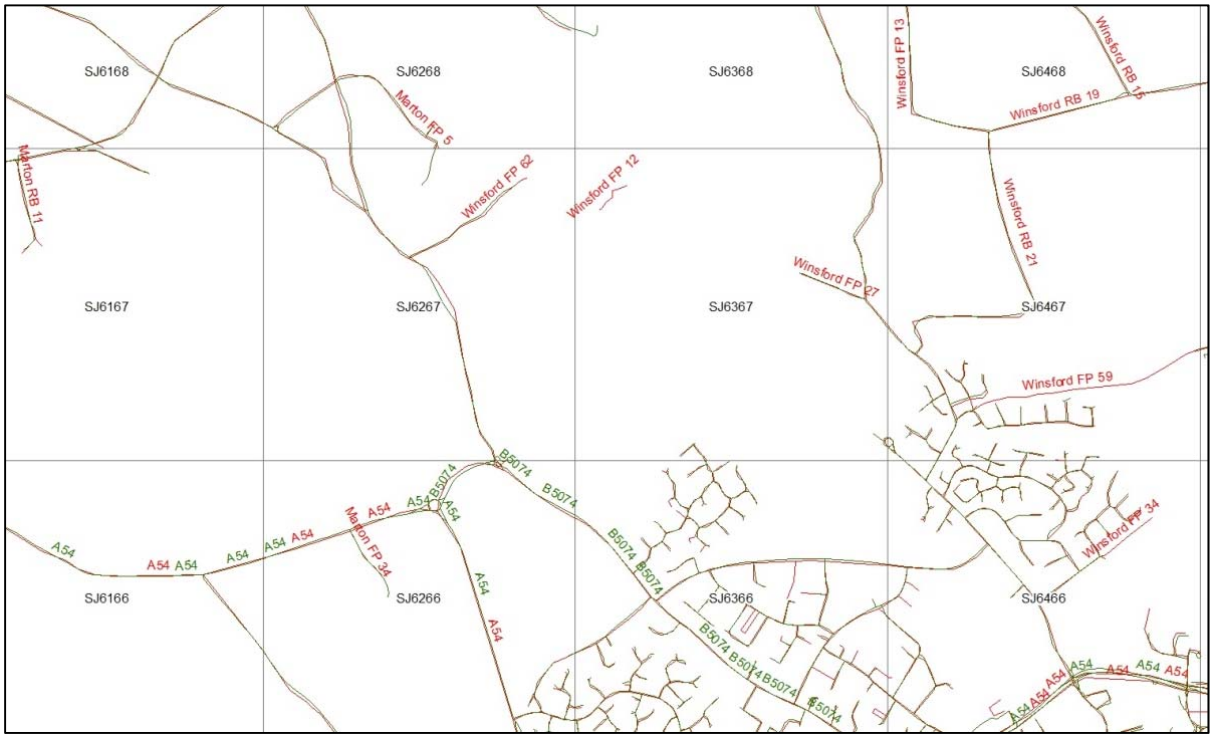

Figure 6.18: *Severn: Secondary names examination finds systematic errors in OSM tagging (red:OSM, green:ITN) (area 1 of Figure 6.10)*

A similar (or maybe opposite) case is in the Isle of Wight (Figure 6.10 – area 2), where an OSM mapping party worked in 2006 and enriched the OSM dataset (OSM wiki pages, 2011). Figure 6.19 shows the secondary names used by both datasets. The ones used by OSM (red) seem to follow the rules about the secondary name, while their format generally agrees with the few ones used by the ITN dataset (green) in the area. However, a better knowledge for the naming scheme of both datasets (for OSM it may require contacting the user) would clarify if this is OSM commission, in other words attributes information that should have also been present in the reference dataset. Another similar case is in the North-East East Anglia region (Figure 6.10 – area 3). No mapping party

is officially recorded for that area, so OSM additional information could be attributed to the diligent work of one or more contributors in the area.

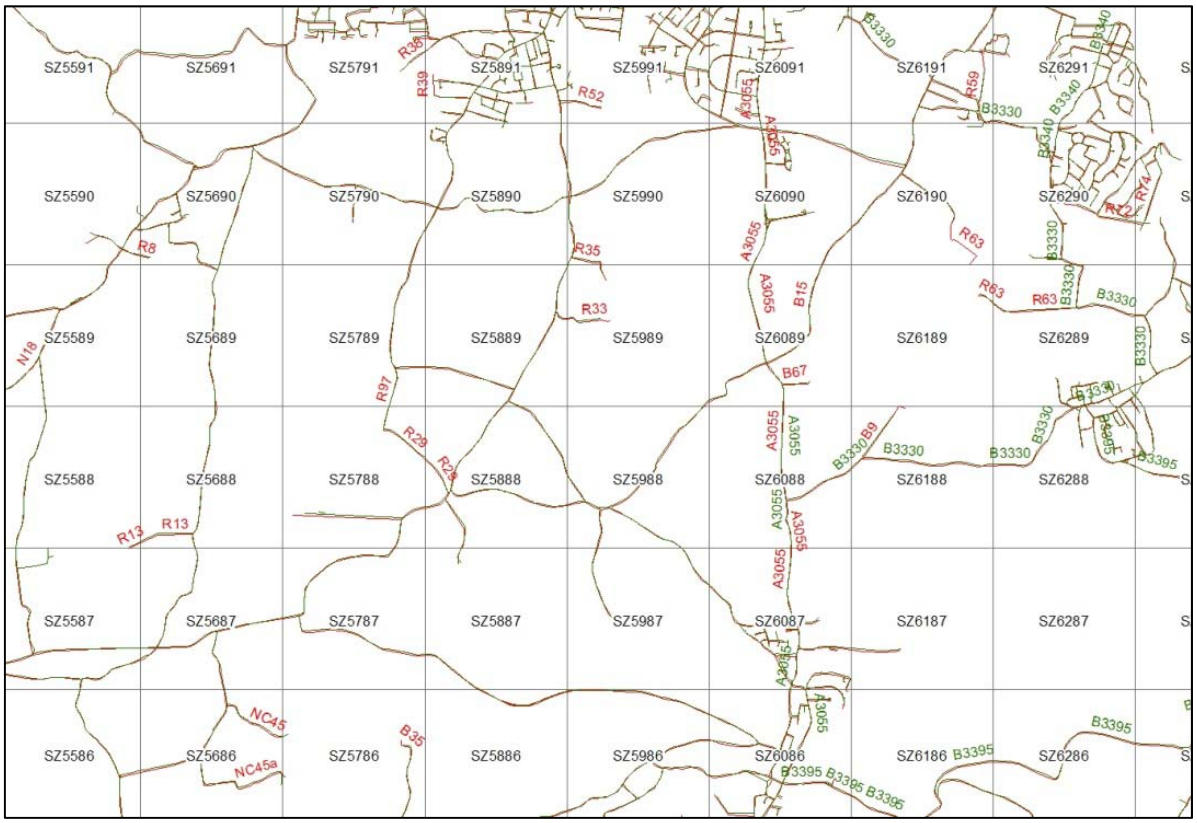

Figure 6.19: *Isle of Wight: Possible VGI commission regarding secondary road names (red:OSM, green:ITN) (area 2 of Figure 6.10)*

6.5.3. VGI commissioned data

A visual examination of the output datasets is necessary to distinguish OSM commissioned data from data matching errors or other types of data that are not specified to be included in the ITN dataset. Figures 6.20 to 6.22 present some examples, showing the ITN dataset in yellow and the OSM commissioned (non-matched) data in red (OSM matched data are not shown). Specifically, Figure 6.20 includes a case of commissioned data and a data matching error, regarding two roads missing from the ITN dataset (north) and a part of an OSM road that failed to be matched (south). Figure 6.21 shows the OSM supremacy in footpaths and trails in Lake District, which however are not described in ITN specifications. Figure 6.22 presents a recently built-up area where OSM seems to be more updated that the ITN dataset.

Table 6.24 provides the tiles where a sampled examination of the OSM non-matched data showed VGI commission, showing in parenthesis the OSM road type (or types) of the commissioned data in

each tile. Some of these examples were already presented or can be found as figures in Appendix C. A more extensive examination could probably provide additional cases of OSM commission.

Figure 6.20: *OSM commission (upper red feature) and data matching error (lower red feature)*

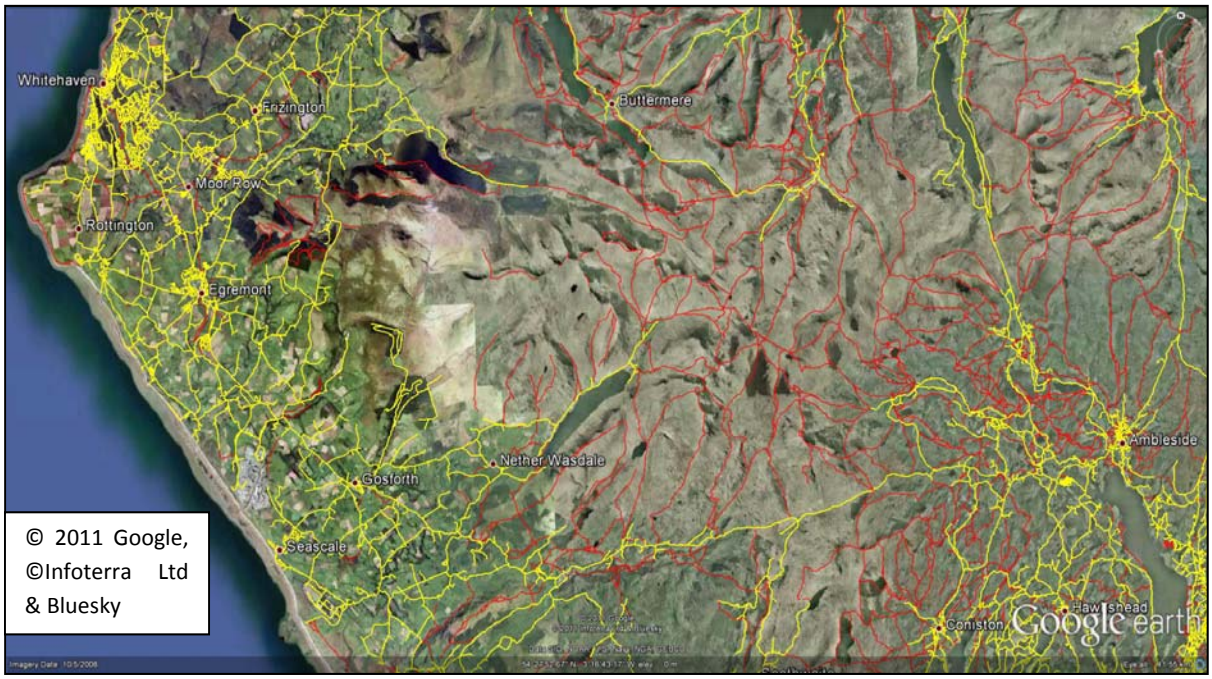

Figure 6.21: *OSM supremacy in Lake District area (paths, footways, tracks, etc)*

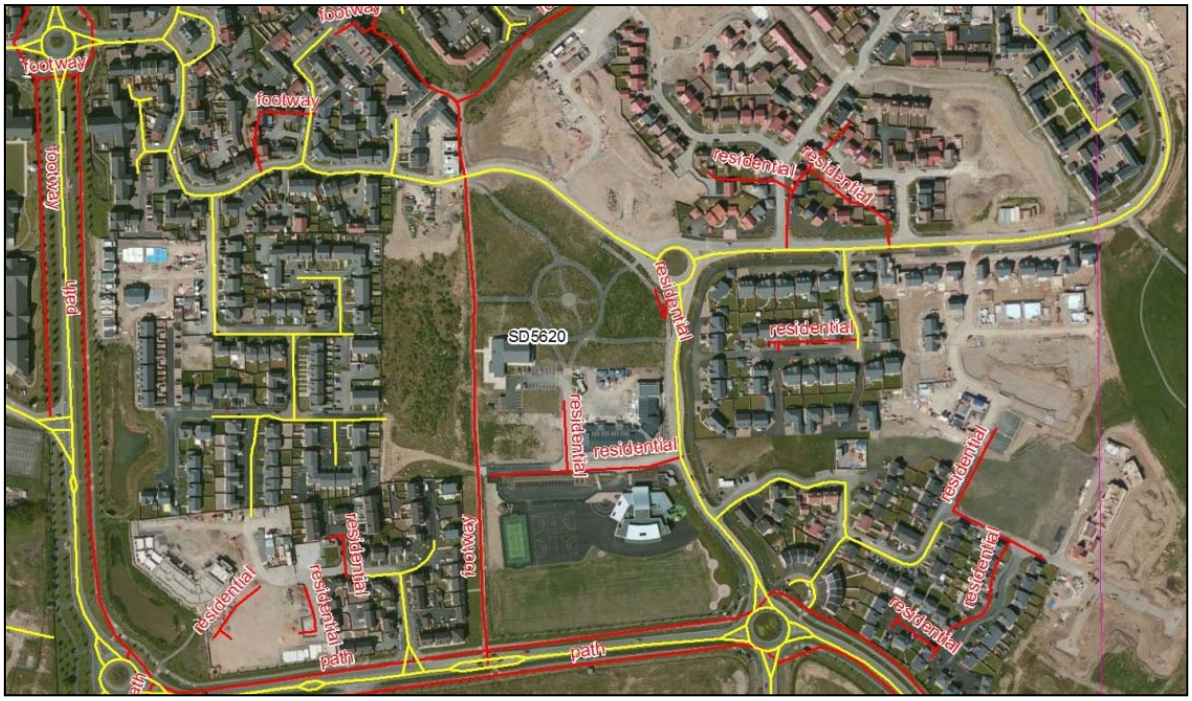

Figure 6.22: *OSM commissioned data in a new built-up area (Satellite imagery source: Bing Maps provided by ESRI's ArcMap)*

Region	Tile	Details	Index
East Anglia	SP6348	(trunk)	
Essex	TL7204	(service)	Apendix C-Fig. 6
Humberside	SK5190	(residential)	Apendix C-Fig. 5
	SK4383	(residential, service)	
Lancashire	SJ3885	(residential)	Apendix C-Fig. 3
	SD5620	(residential): OSM more updated than ITN	Fig. 6.22
London	TQ0976	(service)	Apendix C-Fig. 1
	TQ0879	(residential, service, unclassified)	Apendix C-Fig. 2
	TQ0778	(unclassified): commission and data matching error	Fig. 6.20
	TQ4281	(residential, service)	Fig. 5.27
Manchester	SJ9094	(residential)	
	SD8902	(residential)	
Midlands	SP1791	(residential)	
North	NY6587 & NY6588	(bridleway)	Fig. 5.26
	NY7061	(path, cycleway): Routes for pedestrians	Apendix C-Fig.8
Severn	SJ9339	(residential): OSM false data?	Apendix C-Fig.4
South	SU3811	(residential, service)	
South East	TQ3205 & TQ3206	(residential, service)	
South West	SX9591	(residential)	
Wales	SH6267 & SH6268	(living_street)	
West	SP5822	(service, unclassified)	
Yorkshire	SE1945	(service, track, residential)	Apendix C-Fig. 7

Table 6.24: *Tiles with OSM commission*

6.5.4. Topology

During the manual evaluation, there were a few cases in rural areas where OSM features were not divided when intersected, which resulted in having to compare complex objects that correspond only partially to the other dataset. In such cases it is difficult to decide for the whole feature correspondence even in manual data matching. Even though the automated process follows the rules that generally seem to apply to all data, such cases are difficult to be evaluated as correct or wrong. This is an issue of incorrect topology, already mentioned in section 5.5.4. A way to solve this would be to correct the OSM dataset before the procedure using a spatial function that divides features into smaller ones at road junctions. In this way data matching is likely to improve slightly, as the corresponding part of the initial feature will be addressed differently. In this study case data matching errors remain low despite the sporadic incidents of the above nature (12 features found within the evaluated tiles), however for different datasets with much worse topology, correcting it may be more valuable.

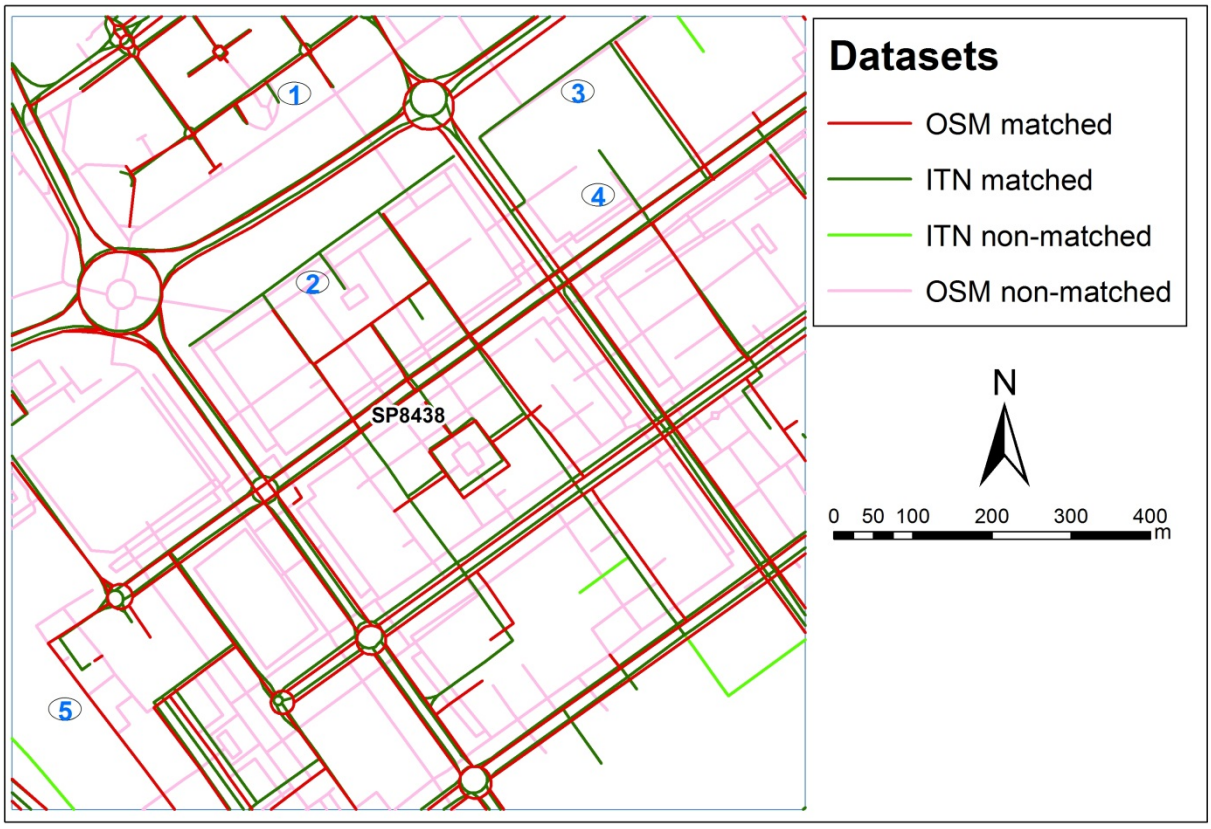

Figure 6.23: *Incorrect VGI topology and its effects on data matching*

Figure 6.23 shows one of the tiles where the problem was most profoundly found and its effects on data matching. Numbers 1 to 4 show non-matched OSM features that could have been matched if they were simpler as shapes, e.g. in case 2 the OSM non-matched feature includes its parallel line to

the south as well as the vertical ones. Case 5 is similar but in the opposite way. A simpler matched OSM feature would only be partially matched (until the road junction next to the right of number 5). Since data completeness relies on matched features' length, such cases affect quality results. Additionally, case 5 may lead to a big buffer value or a lower positional accuracy result, which will underestimate the positional quality of this tile.

6.5.5. Comparison with results from previous studies

Figure 6.24: *Length difference between OSM and OS's Meridian 2 datasets. Black= areas of good OSM coverage; grey=areas of poor OSM coverage (from Haklay, 2010c, p.693)*

Regarding OSM completeness, section 3.4 presented Haklay's (2010c) research for OSM in England, who also used tiles of 1 km^2 to split data and represent heterogeneity. His results (Figure 6.24) refer to 93% of England, leaving out coastal or other 'incomplete' tiles with an area less than 1 km^2. The reference dataset was the 'generalised and incomplete' (p. 692) 'Meridian 2' dataset from OS. The comparison did not include any data matching procedure, but relied on length comparison between the datasets for each tile. Although some OSM road types were removed as non-compatible with Meridian 2 specifications, leading to a comparison of more homologous datasets, he mentioned that this is a matter of attribute accuracy, and it is assumed that OSM road type classification is correct. This was the first approach for OSM accuracy on a national level, although quite rough: tiles with OSM network length equal to or bigger than the Meridian length were considered to have good OSM coverage and vice-versa. A spatial pattern mentioned by Haklay (2010c), is the rectangular area of increased OSM coverage around London, (Figure 6.24) which is where high-resolution satellite imagery was available to OSM mappers.

This thesis provides a more systematic approach, which examines all data and uses the most detailed official dataset available, including a data matching algorithm that provides more accurate completeness results and integrates also the evaluation of additional quality elements. Although Haklay's (2010c) method and limitations allows for a representation of data completeness results using only two classes (referring to which dataset has better coverage), Figure 6.24 (Haklay's results) bears some similarities with the much more detailed results of Figure 6.3 (ITN matched percentages – OSM completeness).

6.6. Summary

For the second case study described in this chapter, the automated method is applied on a national level, using areas of mixed network density. The chapter structure slightly differs from the other case studies due to this mixed type of data. Specifically, in contrast to the other two case studies, there is no use in calculating the average length per tile for each region and dataset (Table 6.1 as compared to Tables 5.1, 7.1), due to the network density variations. For the same reason, evaluation of the stages contribution to data matching and road type correspondence examination are not discussed, as this will not provide any useful information for the whole area (as opposed to sections 5.4.1, 5.5.2 and 7.4.1, 7.5.2 respectively). However, they are both calculated for each region as part of the automated process.

On the other hand, due to the complex network density and larger area, VGI evaluation extends the limits of a well-mapped area or an area of a specific group of users, enabling the realisation of spatial patterns. Areas of different VGI coverage can be visually identified. Areas of similar data completeness, attribute or positional accuracy can also be easily found. Massive errors in VGI due to one or more users not following OSM rules regarding the way of tagging can be distinguished. However, spatial patterns do not necessarily follow the regional borders used in this case study. This is another consequence of VGI heterogeneity. Results also show that a correlation between quality elements is not obvious, for example an area can be quite complete in data or attributes but with low positional accuracy.

The evaluation showed that OSM includes data with incorrect topology. Although this was also noticed in rural areas of the previous case study, corresponding results showed that data matching is not significantly affected. In this study case, however, their influence is stronger, as most of the examined network is in rural areas, and affects positional accuracy results by producing an increased number of outliers. Although data matching is also affected, data matching errors remain quite low. In cases of VGI sources with incorrect topology, the findings of this case study suggest that some data preparation, regarding splitting VGI features into new ones when intersected, will be essential for the automated method to reduce data matching errors and provide more accurate quality results.

OSM proves to have additional data in many areas throughout the country, however, they mostly refer to data not collected by ITN (footpaths, cycleways, etc). Manual examination of the non-matched data marked with commission indication can show cases of OSM being more complete than ITN. Combined with the OSM road type attribute, data not present in the ITN dataset can be isolated to be used for a specific purpose (e.g. enhance the ITN dataset with service road types, usually representing parking areas or other roads that are publicly accessible and suitable for a car).

Despite the valuable findings of this chapter, data sources remain the same with the previous study case, so data structures between the compared datasets remain unchanged. The next chapter moves further and applies the automated method on different data sources and areas, to ensure its robustness and efficiency and upgrade it to a framework for VGI quality evaluation in general.

Chapter 7

Third Case Study: Haiti (Port-au-Prince)

7. Third case study: Haiti (Port-au-Prince)

7.1. Introduction

The method proposed (Chapter 4) was tested for its efficiency in areas of different but relatively homogeneous network density and accuracy (Chapter 5), as well as in larger and heterogeneous areas (Chapter 6). Although it proved to work efficiently even on a national level, it is required to examine if it is generic and applicable to other VGI linear datasets as well. This will ensure that the developed framework is of maximum value to end-users. Hence, this chapter applies the framework on Haiti, using the U.N. Stabilization Mission for Haiti ('MINUSTAH') project as reference source and Google Map Maker (GMM) as crowd-sourced one, although OSM is also available in the area. The analysis follows a structure similar to the previous two chapters, starting with area justification and data preparation, applying the method according to the flow diagram of Figure 4.1, presenting the results, evaluating the method and concluding with discussion.

Section 2.3 presented the crowd-sourced projects OSM and GMM that will be used in this case study. The next section briefly describes the UN dataset, which will be used as a reference data source.

7.1.1. Reference Data Sources: United Nations' 'MINUSTAH' dataset for Haiti

The United Nations Stabilization Mission in Haiti, called 'MINUSTAH', started in 2004 (U.N., 2011). After the disastrous earthquake of 2010, U.N. released spatial information in order to be used by the rescue teams in the area. The road network is described in three shapefiles. The first one includes the major roads for the whole country. The second one is richer by also including minor roads for the whole country, however major roads between the two datasets do not match (Figure 7.1a). The third one covers Port-au-Prince in much more detail (Figure 7.1b), which makes it the most detailed official dataset. Roads are classified in 5 groups, while there are also objects with null road type value. This dataset is selected to be used in this case study.

Figure 7.1: *Incompatibility between UN datasets in the area of Port-au-Prince,* ***a:*** *'hti_rdsl1' (major roads) and 'hti_rdsl2' (major and minor roads)* ***b:*** *'hti_rdsl1' and 'portauprince_rdsl'*

7.2. Area justification and data preparation

After the earthquake in 2010, apart from the official spatial datasets released to be used by the rescue teams, crowd-sourced spatial information was gathered by volunteers for web mapping applications for the same purpose. It is an example of immediate response of VGI to natural disasters (Mullins, 2010; Haklay, 2010b) and it is among the few areas where there is free access to official data (from U.N.), VGI (from OSM) and proprietary but still crowd-sourced data (from Google Map Maker). This enables:

- testing the method on completely different data sources than the ones used so far.
- testing the method by comparing two different web-mapping projects based on volunteers, in other words comparing VGI with VGI, which adds a new dimension to this research.

From the VGI side, the relevant datasets from Google Map Maker (GMM) were downloaded, among which there is the shapefile selected in previous section. The data refer to the whole country. For OSM, however, there are two versions covering the area. One is from the geofabrik website (geofabrik, 2010), which appears to be the same as the U.N.'s 'MINUSTAH' shapefile. Specifically, although there are some variations in the attributes, objects are the same as in the U.N. file with exactly the same geometry. The second version is from the planetdump website (planetdump, 2010), which appears to be different from U.N.'s dataset and will be used here. This shapefile also refers to the whole country.

All datasets were downloaded in March 2011. They are all in WGS84 Geographic Coordinate System and they had to be re-projected to UTM WGS84 Zone 18N (EPSG 32618), so that measurement units would refer to meters instead of degrees. This is essential for the method to work, as all equations described in Chapter 4 use such metric system. Following a tile size of 1 km^2 (see discussion in section 4.16.1), a grid was created using Manifold GIS software for the whole country. The tiles covering the U.N. dataset for Port-au-Prince were selected and exported to be used as the tessellation file. Using the outer boundaries of this sub-grid, VGI datasets (GMM and OSM) were clipped and then they were loaded in the PostGIS database, along with the tessellation file.

Section 5.5.4 argued that topology was not necessary to be examined for the first case study, while section 6.5.4 noticed an increased importance in its examination due to some errors found in VGI topology. Since this had a minor influence on data matching and quality results of the second case study, it was not corrected. However, in Haiti case the topological quality of both reference and VGI datasets is significantly lower than in previous case studies (see Figure 7.17). Following a structure

similar to the previous chapters, the topology issue is further discussed in section 7.5.4. In this section it is suffice to say that in order to deal with this problem, all datasets had to be pre-processed so that all features would be divided at road intersections. A new serial ID is created to be used as primary key for the new objects of each dataset, since by further dividing the features their ID is no longer unique.

Another variation of the datasets, compared to the previous study cases, is their lack of road name attributes in general. There is no secondary name attribute in either dataset, while UN has a primary name attribute only for 29% of the road network. For the rest 71% UN's primary name has a value of 'NoName' or 'No name' instead of null, which had to be removed, otherwise it would be treated as a road name value. GMM, on the other hand, has primary name information only for 52 features (0.31% of the total network length). This lack of attributes poses another challenge: how efficient would the proposed method be when attributes barely exist in one dataset?

Comparison in this case study firstly involves U.N. (reference) with Google Map Maker (VGI) (Figure 7.2a). Section 2.3 provided a brief description of Google Map Maker (GMM) data source. However, the existence of two crowd-sourced data sources in the area enables their comparison as well, included in this chapter as a second sub-case, where GMM is considered as reference dataset and OSM as VGI (Figure 7.2b). There are two reasons for selecting GMM as reference dataset. Firstly, GMM road types are strictly structured in classes, similarly to the ITN dataset used as reference in previous chapters, which enables the examination of road type correspondence. Secondly, OSM seems to be by far the richest dataset in the area (see Figure 7.2b and Table 7.1, which presents the network lengths and area size, similarly to Table 5.1), so it would provide many examples of VGI commission and more opportunities to evaluate the method performance.

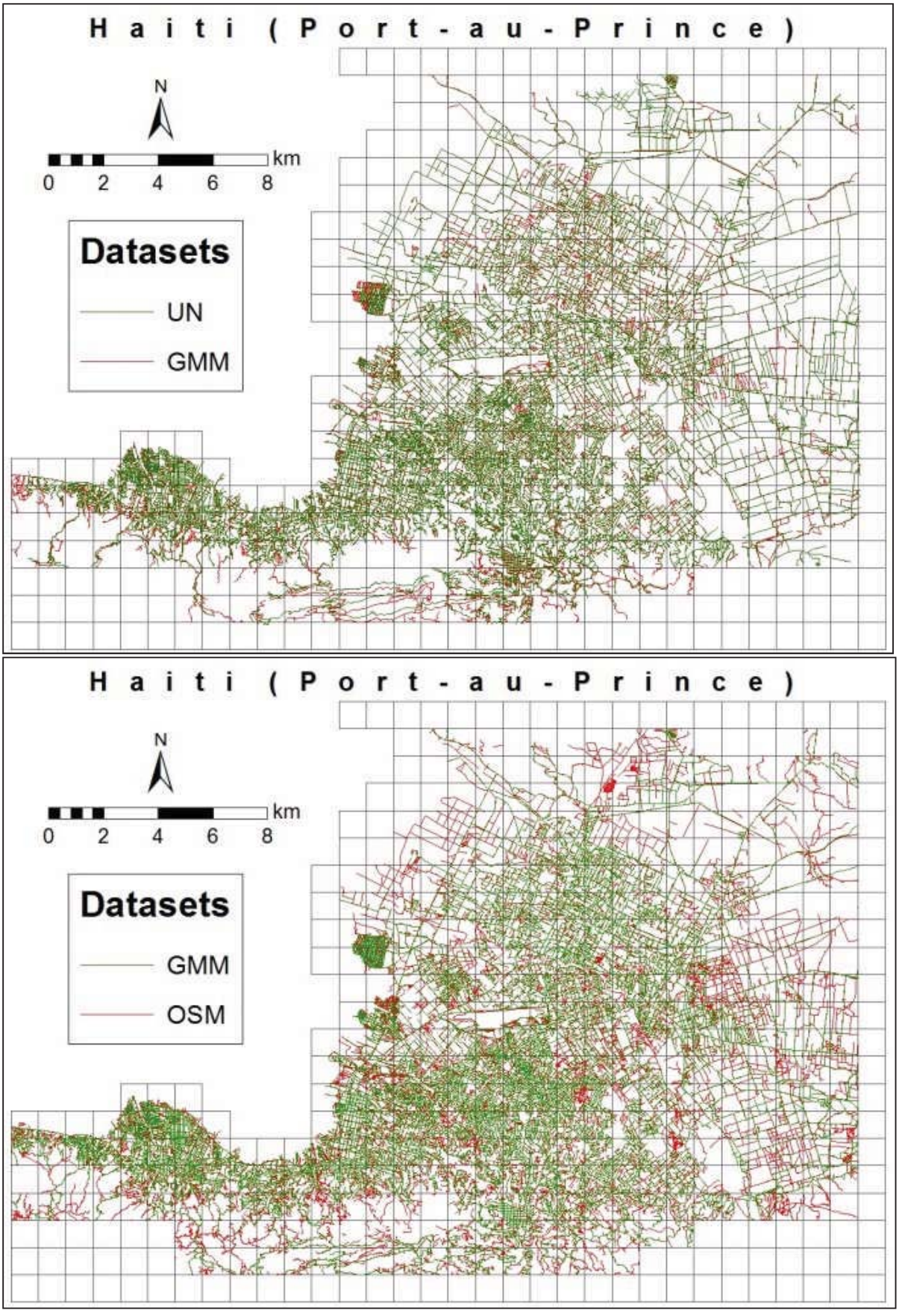

Figure 7.2: *Area and datasets studied in the 3rd case study*

Area	Total Area size	Dataset	Total network length (m)	Average length (m) per tile
Haiti (Port-au-Prince)	409 km^2	UN	2,516,730	6,153
		GMM	2,385,926	5,834
		OSM	3,902,001	9,540

Table 7.1: *Studied areas and road network information for 3rd case study*

7.3. Results

Results in this chapter are produced following the same method (see flow diagram of Figure 4.1) as in previous chapters, with no further modifications, which prove the generality of the proposed methodology.

7.3.1. UN (reference) and GMM (VGI) comparison

Figures 7.3 and 7.4 present the output matched and non-matched datasets respectively. Table 7.2 presents the total lengths (similarly to Table 5.2 and its relevant description). (Detailed CSV files that describe each tile individually are produced along with the relevant shapefiles).

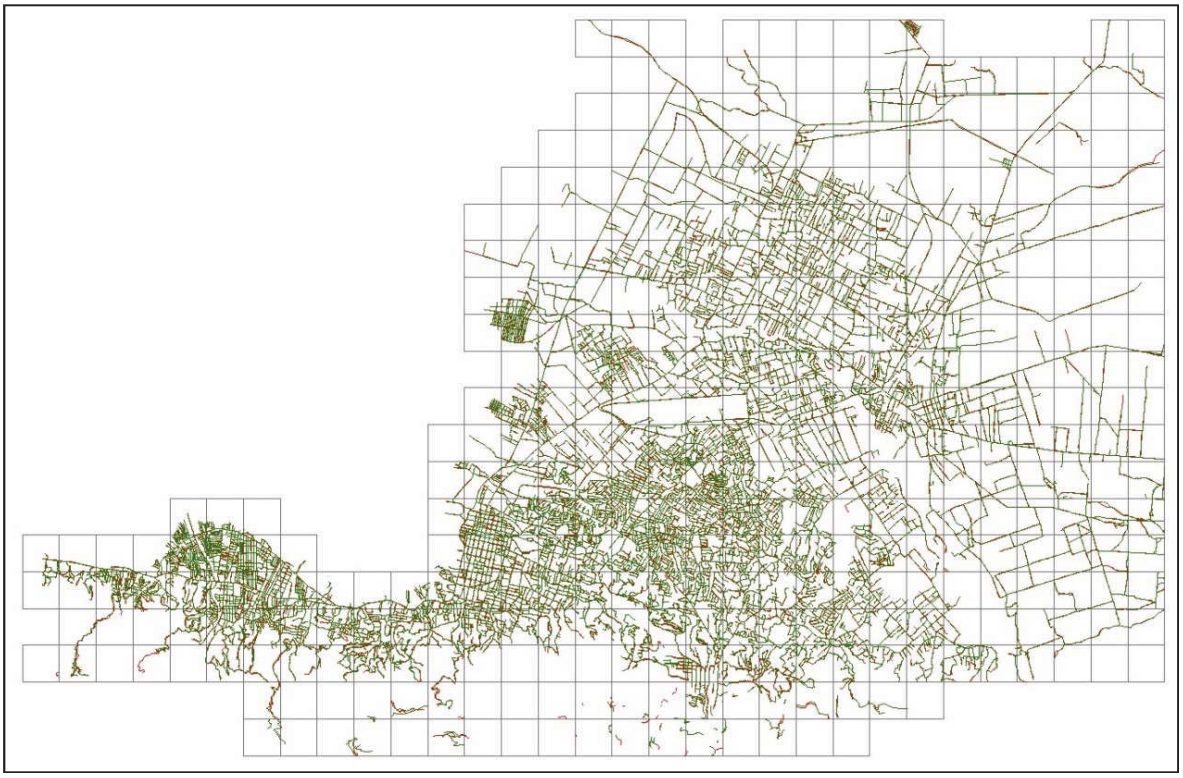

Figure 7.3: *Matched reference (UN - green) and VGI (GMM - red) dataset*

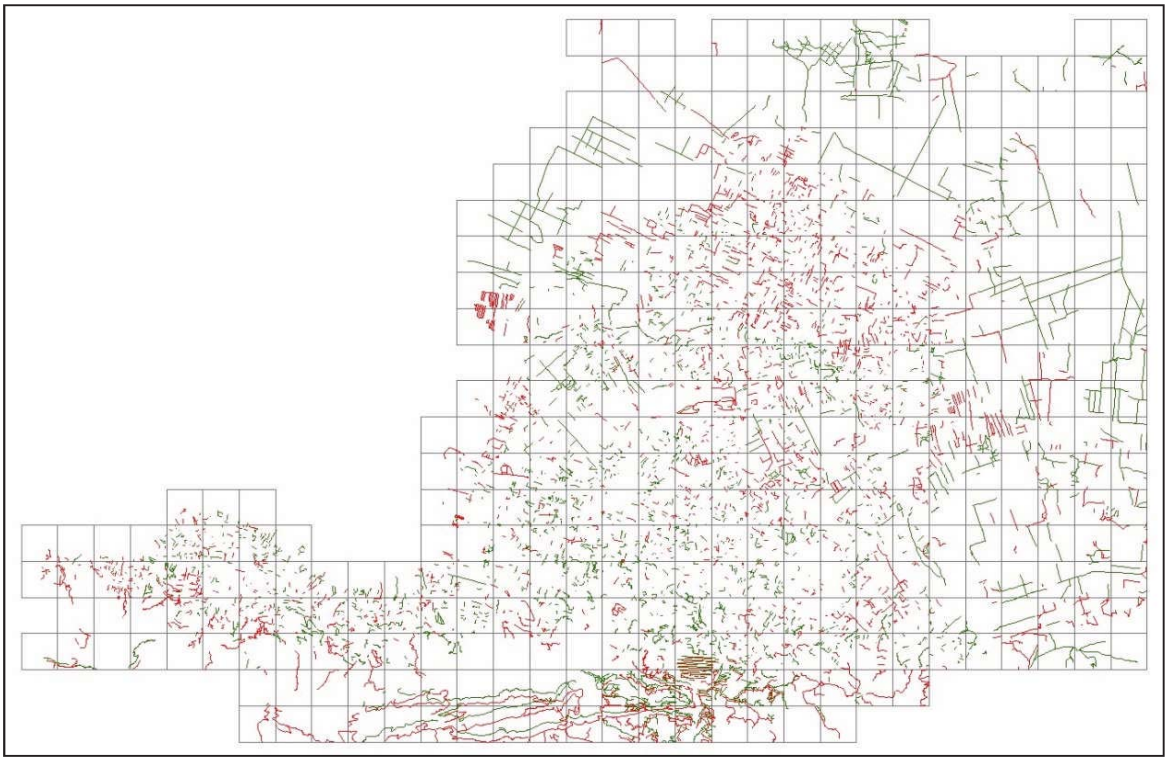

Figure 7.4: *Non-matched reference (UN – green) and VGI (GMM - red) dataset*

Area	Dataset	Total Length	Length (m) Compared	Length (m) Matched	Length (m) non-matched
Haiti	UN	2,512,888 m	2,490,839 (99.12 %)	1,998,098 (79.51 %)	516,790 (20.49%)
	GMM	2,385,926 m	2,373,299 (99.85 %)	1,899,106 (79.90 %)	477,760 (20.10%)

Table 7.2: *Resulting UN and GMM network lengths for Haiti area*

Figure 7.5 presents data completeness and positional accuracy, while Figure 7.6 refers to attribute accuracy (only primary name is present). The classification used is the same as in the first case study, already justified in section 5.3, along with the meaning of blank tiles.

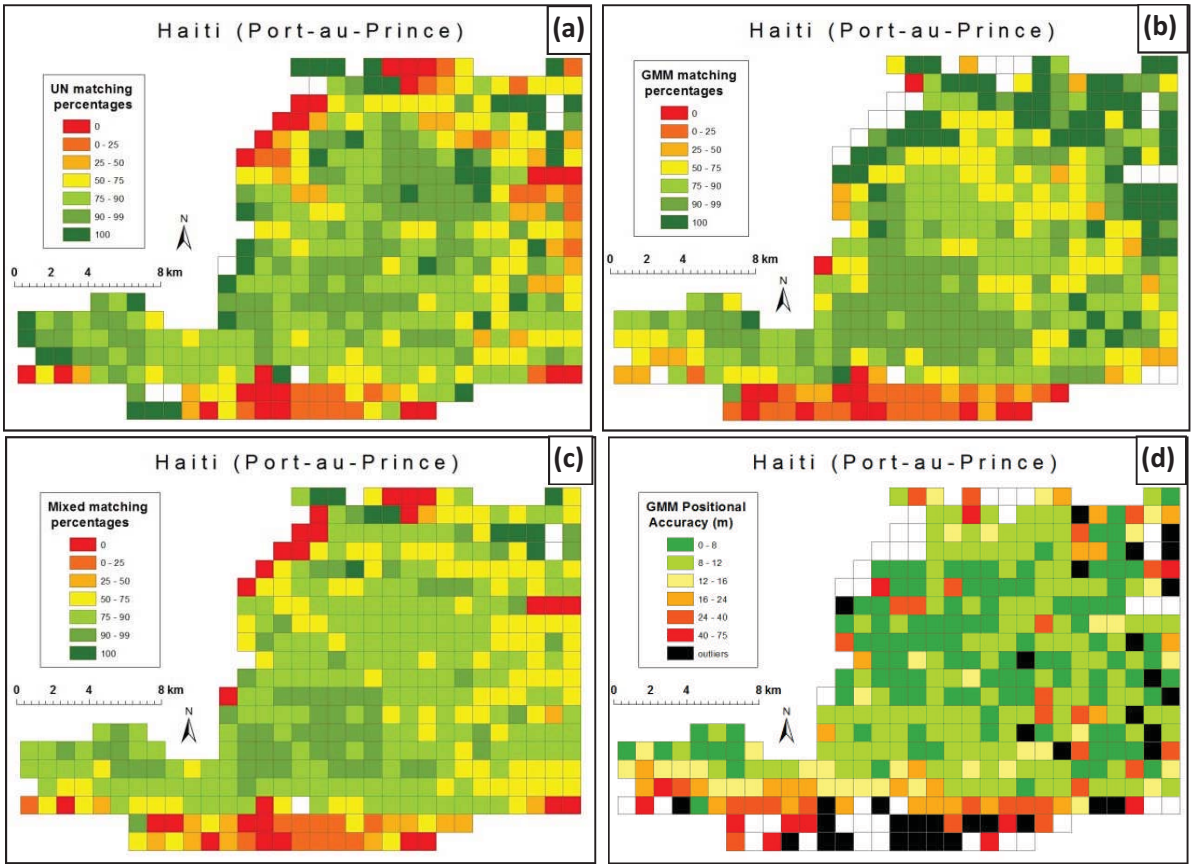

Figure 7.5: *Haiti area – data completeness and positional accuracy,* ***a:*** *UN matching percentages (GMM completeness)* ***b:*** *GMM matching percentages (GMM commission)* ***c:*** *Mixed percentages (level of agreement between datasets)* ***d:*** *Positional accuracy*

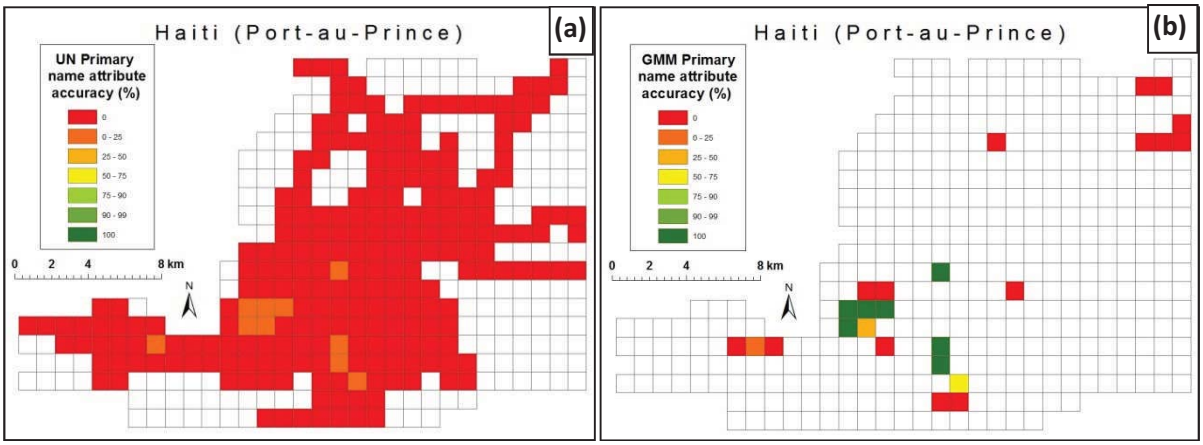

Figure 7.6: *Haiti area – primary name,* ***a:*** *UN percentages (GMM attribute accuracy)* ***b:*** *GMM percentages*

Figure 7.5a shows that GMM is more complete in the city centre, demonstrated by UN matching percentages mostly above 75%, as opposed to the suburbs. Additionally, Figure 7.5b shows that GMM also contains additional features not present in the reference dataset, with GMM matching

percentages generally less than 75%. Positional accuracy is quite good in the city centre (less than 12 or even 8 meters), but deteriorates when moving to the suburbs. This is also the reason why 26 tiles considered as outliers (6.4% of total area) are mostly gathered there. Finally, Figures 7.6a and 7.6b show that UN is much richer than GMM in attributes, while Figure 7.6b shows that even for the few existing road names in the VGI dataset, attributes are not so compatible with the reference dataset.

Table 7.3 provides more statistics regarding the results distribution (their meaning was discussed in section 5.3), based on the results for the tiles with the available data for each evaluation ('Tiles compared' column).

Urban area (Greater London)	Dataset	Tiles compared	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	UN	402	71.32	82.51	-1.25	92.25	29.30
	GMM	390	75.90	83.52	-1.52	95.64	26.69
Primary name accuracy	UN	237	0.14	0.00	5.86	0.00	0.74
	GMM	25	33.24	0.00	0.77	100.00	45.39
Positional accuracy (outliers ignored)	GMM	336	13.59	9.25	2.61	14.38	10.91

Table 7.3: *Statistics for Haiti area (UN and GMM datasets)*

7.3.2. GMM (reference) and OSM (VGI) comparison

Figures 7.7 and 7.8 present the output matched and non-matched datasets respectively. Table 7.4 presents the total lengths (similarly to Table 5.2). (Detailed CSV files that describe each tile individually are produced along with the relevant shapefiles).

Area	Dataset	Total Length	Length (m) Compared	Length (m) Matched	Length (m) non-matched
Haiti	GMM	2,385,926 m	2,385,926 (100 %)	2,292,638 (96.09 %)	93,288 (3.91%)
	OSM	3,902,001 m	3,863,723 (99.02 %)	2,331,739 (59.76 %)	1,570,262 (40.24%)

Table 7.4: *Resulting GMM and OSM network lengths for Haiti area*

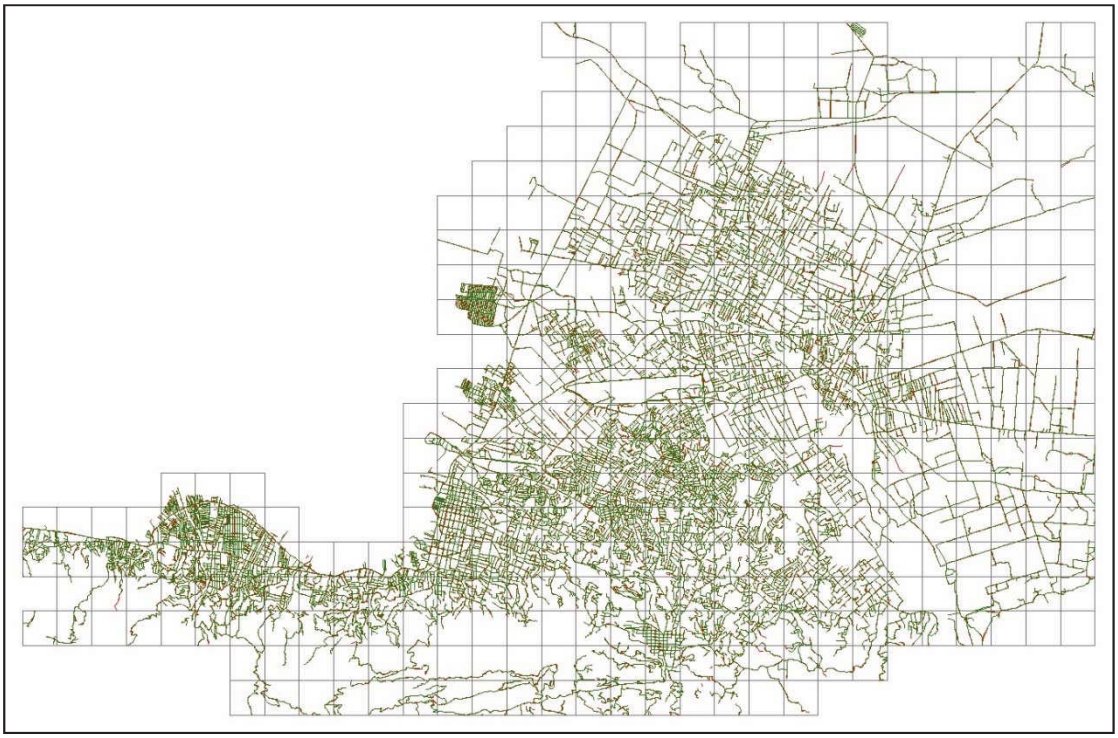

Figure 7.7: *Matched reference (GMM - green) and VGI (OSM - red) dataset*

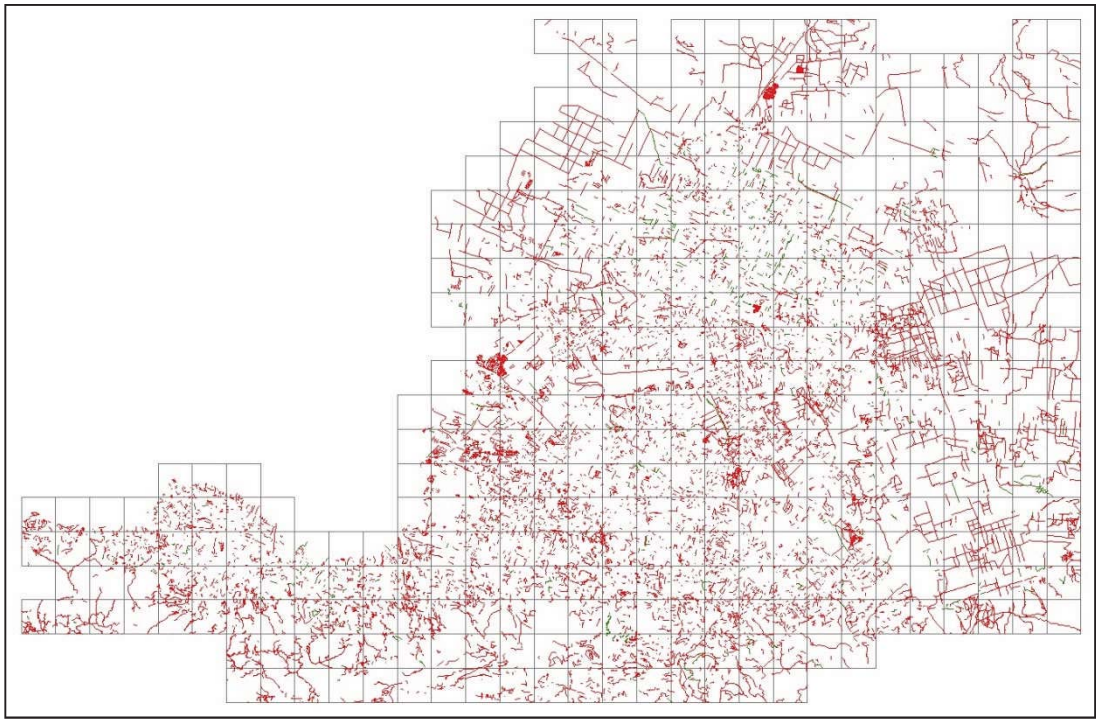

Figure 7.8: *Non-matched reference (GMM - green) and VGI (OSM - red) dataset*

Figure 7.9 presents data completeness and positional accuracy, while Figure 7.10 refers to attribute accuracy (only primary name is present). The same classification as the one in previous section is used to enable visual comparison.

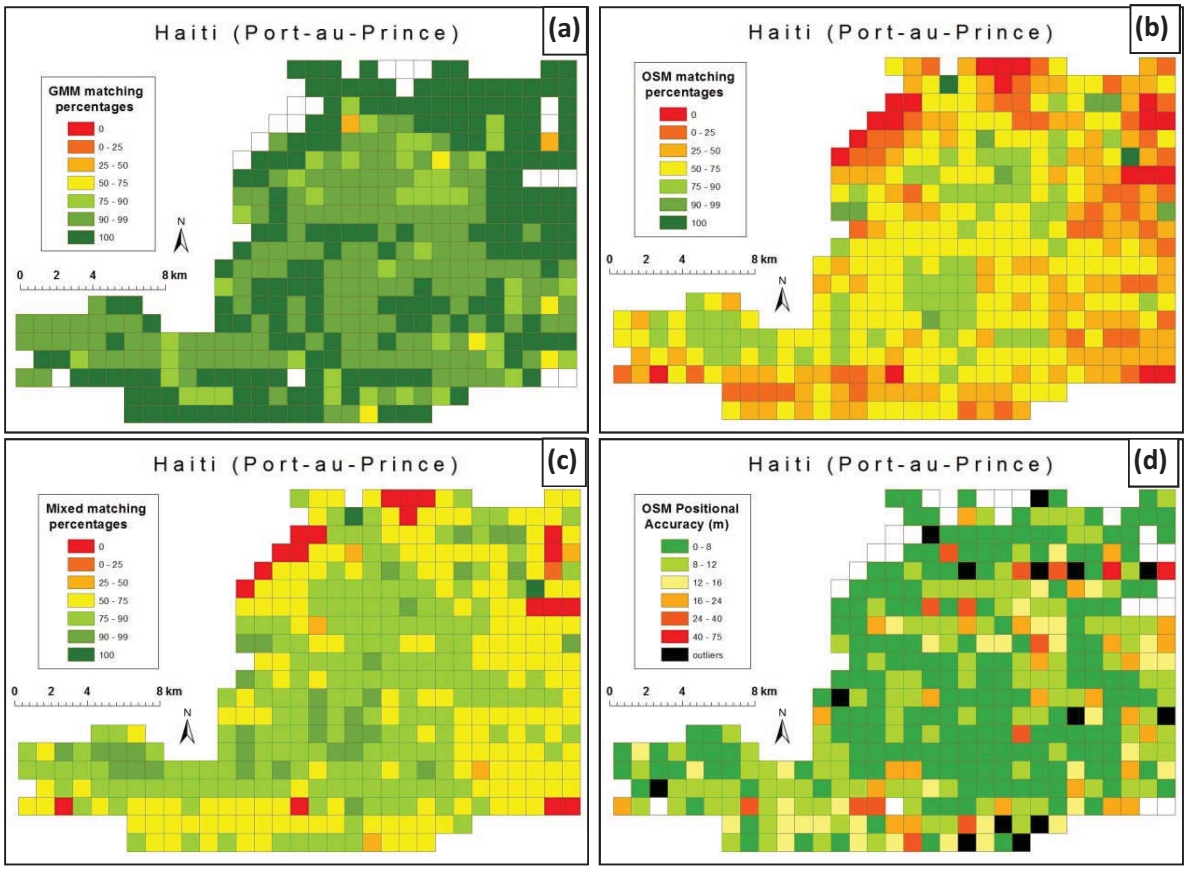

Figure 7.9: *Haiti area – data completeness and positional accuracy,* ***a:*** *GMM matching percentages (OSM completeness)* ***b:*** *OSM matching percentages (OSM commission)* ***c:*** *Mixed percentages (level of agreement between datasets)* ***d:*** *Positional accuracy*

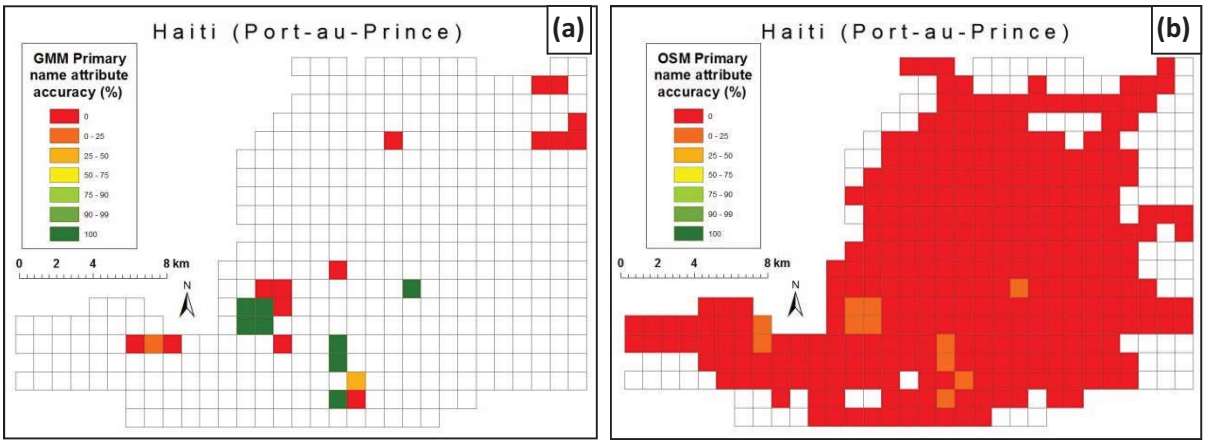

Figure 7.10: *Haiti area – primary name,* ***a:*** *GMM percentages (OSM attribute accuracy)* ***b:*** *OSM percentages*

Figure 7.9a shows that OSM is more complete in the city suburbs, which is demonstrated by UN matching percentages mostly at 100% and GMM matching percentages generally less than 50%. Positional accuracy is quite good for the whole area (most tiles score less than 8 meters). Finally,

Figures 7.10a and 7.10b show that OSM is much richer than GMM in attributes, including road names in most of the tiles (Figure 7.10b) as opposed to only 25 tiles for the GMM dataset.

Table 7.5 provides more statistics regarding the results distribution (their meaning is discussed in section 5.3), based on the results for the tiles with the available data for each evaluation.

Urban area (Greater London)	Dataset	Tiles compared	Average pct	Median pct	Skewness of pct	Quartile 3 of pct	St.Dev of pct
Data matching (completeness)	GMM	390	96.00	99.52	-3.88	100.00	7.63
	OSM	409	52.31	55.89	-0.46	71.42	24.56
Primary name accuracy	GMM	25	34.04	0.00	0.76	100.00	46.70
	OSM	318	0.09	0.00	9.93	0.00	0.63
Positional accuracy (outliers ignored)	OSM	373	9.87	8.13	3.50	11.08	6.95

Table 7.5: *Statistics for Haiti area (GMM and OSM datasets)*

7.4. Evaluation

7.4.1. Contribution of stages

Table 7.6 provides information on the contribution of each stage to the matching procedure for the two cases examined (between UN and GMM as case 1 and between GMM and OSM as case 2). As expected, stages 2 and 3 do not participate much due to the lack of road name attributes in one dataset of each case, and the most significant is stage 4, similarly to the rural area of Chapter 5. However the area type here is mainly urban, which is also demonstrated by the relatively low efficiency of stage 1: in dense networks it is harder to achieve 1-1 matching of objects with no other possible candidates.

Dataset	Area	Matched percentages (%) compared to the total matched length					
		Stage 1	Stage 2	Stage 3	Stage 4	Stage 5-	Stage 5+, 6, 7
UN	Haiti, case 1	10.00	0.07	0.04	89.76	0.00	0.13
GMM		13.83	0.04	0.02	72.47	-0.01	13.64
GMM	Haiti, case 2	14.77	0.04	0.03	84.56	0.00	0.60
OSM		19.71	0.06	0.06	69.37	0.00	10.79

Table 7.6: *Contribution of stages to data matching for the two Haiti cases (UN-GMM and GMM-OSM)*

What is also worth noticing is that stages 5, 6 and 7 are much more important for the overall feature matching than in Chapter 5. The lower-quality topology issue (mentioned in section 7.2 and further discussed in section 7.5.4) could not have been solved by dealing only with segments and skipping stages 5 to 7, because, as Table 7.6 shows, the matching process would not have been so efficient. Stage '5-' and its negative value is justified in section 5.4.1.

7.4.2. Object matching efficiency

A manual evaluation was performed for 40 tiles (about 10%) for both cases, randomly selected (Figure 7.11). Section 5.4.2 described how data matching is manually evaluated. Table 7.7 presents the results for both cases. Data matching errors remain low even in this case study that uses different datasets than in previous studies, with almost no road name attributes, which proves the efficiency and robustness of the method. Figures 7.12 (a and b) are graphs of the total error compared to the number of tiles evaluated, which also shows that 10% is a more-than-adequate percentage to estimate the errors for the whole dataset. Data matching errors are less when comparing the two crowd-sourced datasets (GMM and OSM). The reason is than in the first case (UN and GMM) there are many examples of objects that although they seem to be corresponding ones, they fail to be matched due to their increased distance, which exceeds the thresholds used to define the objects correspondence (see section 4.8). Such cases are visible in the southern tiles of Figure 7.3, while the same problem does not seem to apply to the second case of GMM and OSM datasets (Figure 7.8). This is further discussed in section 7.5.3, while Figure 7.14 further provides a closer look on this reference data shift.

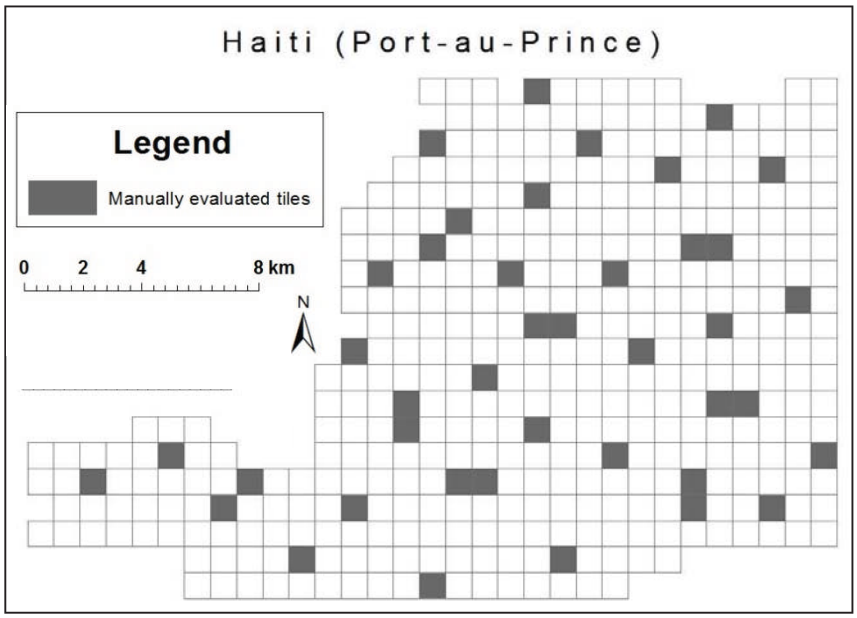

Figure 7.11: *Randomly selected tiles for manual evaluation*

Data-set	*Area*	*Length (m) compared*	*Length (m) evaluated*	*Missing length (m)*	*Surplus length (m)*	*Total matching error (m)*
UN	Haiti, case 1	2,490,839	298,193 (11.97%)	3,701 (1.24%)	6,404 (2.15%)	**10,105 (3.39%)**
GMM		2,373,299	288,675 (12.16%)	6,134 (2.13%)	850 (0.29%)	**6,984 (2.42%)**
GMM	Haiti, case 2	2,385,926	288,675 (12.10%)	662 (0.23%)	2,539 (0.88%)	**3,201 (1.11%)**
OSM		3,863,723	459,362 (11.89%)	4,589 (1.00%)	2,270 (0.49%)	**6,859 (1.49%)**

Table 7.7: *Data matching errors between UN, GMM and OSM datasets*

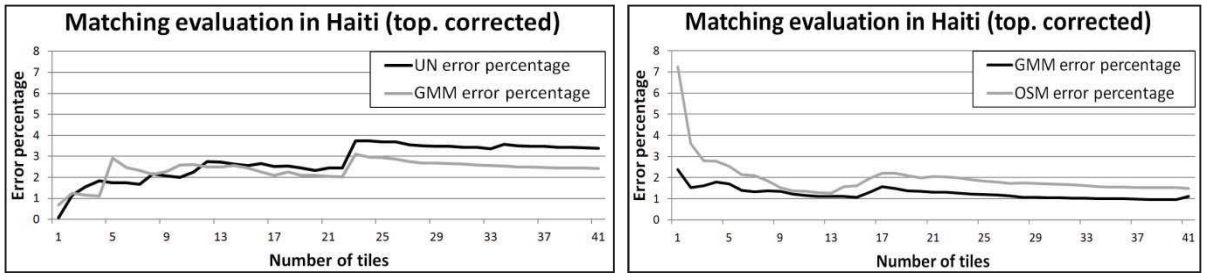

Figure 7.12: *Data matching error levels between* ***a:*** *UN and GMM,* ***b:*** *GMM and OSM datasets*

7.4.3. Attribute accuracy efficiency

Section 5.4.3 described how attribute accuracy is manually evaluated, generally distinguishing three types of error. Due to the low level of attribute information in the GMM dataset, all GMM data with name attributes were examined (the tiles described in Figure 7.6b or 7.10a). Results (Table 7.8) show a reduced efficiency of the attribute accuracy estimation (error levels 9.61% and 7.62% for UN-GMM and GMM-OSM comparisons respectively). Some examples in Table 7.9 show that the applied text similarity thresholds cannot cover more significant inconsistencies between road names.

Haiti area (Port-au-Prince)	**1st case: UN and GMM comparison**				**2nd case: GMM and OSM comparison**			
	UN matched		**GMM matched**		**GMM matched**		**OSM matched**	
	Length (m)	**Pct (%)**	**Length (m)**	**Pct (%)**	**Length (m)**	**Pct (%)**	**Length (m)**	**Pct (%)**
Primary name	146,707	100	6,752	100	6,575	100	215,755	100
Error type 1	1,204	0.82	0	0.00	0	0.00	1,127	0.52
Error type 2	0	0.00	0	0.00	0	0.00	0	0.00
Error type 3	704	0.48	649	9.61	501	7.62	440	0.20
Total errors	1,908	1.30	649	**9.61**	501	**7.62**	1,567	0.73

Table 7.8: *Attribute accuracy errors*

UN Road Names	GMM Road Names	OSM Road Names
BD JJ LA DESALINE	Blvd Jean Jacquea Dessaline	Nationale No 2
RUE LOUVERTURE (1 WAY - EAST)	***Louverture***	***Rue Louverture***
R. S. VINCENT	***Stenio Vincent***	***Rue Stenio Vricent***
-	VILAIRE	Rue E. Vilaire
-	KILLICK	Rue A. Killick
-	AVENIDA MAIS GATE	Ave MaΓƒB—s GatΓƒB©

Table 7.9: *Text similarity failure and success (the latter underlined in bold-italic) in matching strings*

7.4.4. Positional accuracy efficiency

Similarly to what was described in section 5.4.4, the tiles considered as outliers in terms of positional accuracy were manually examined. Table 7.10 provides the results. Data matching errors come first as a reason for outliers, suggesting that it is a useful quality measure. Inconsistencies between the shape and length of the features that are used to describe the real world objects result in increased buffer values (low positional accuracy), which is the second most common case for outliers. Some examples are the ones highlighted in Figure 7.13 with a yellow ellipse. These are cases of 'simple different representation' of objects, which, as section 6.4.3 discussed, cannot be predicted or topologically corrected. Finally, outliers due to bigger distances between corresponding objects come last. Similarly to the previous case studies (Tables 5.9 and 6.20), these low values suggest that the thresholds used in section 4.12.4 to define the positional accuracy outliers are appropriately selected.

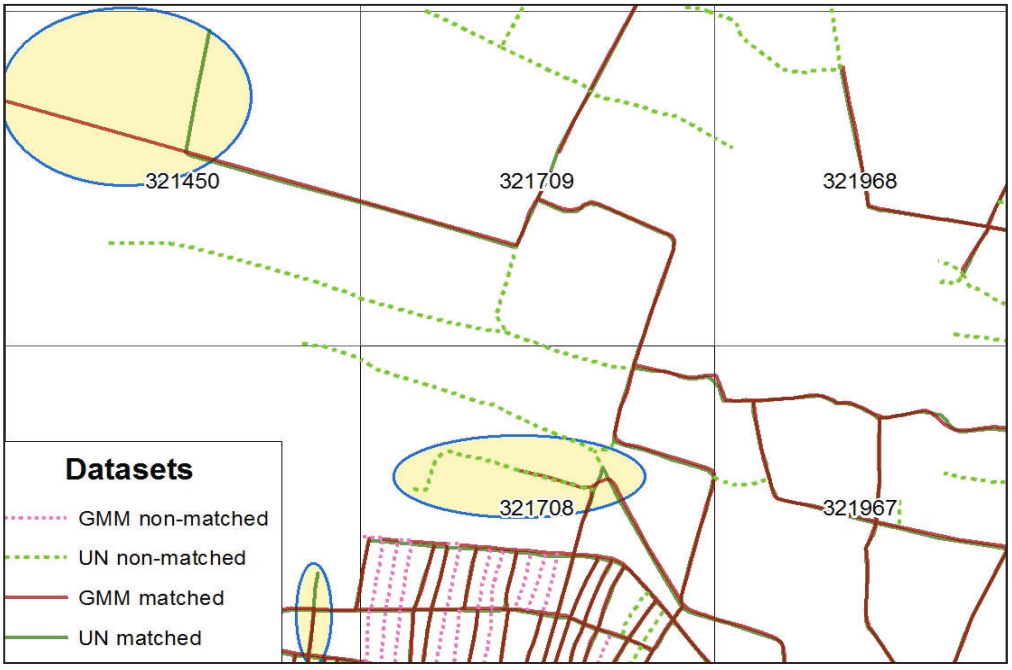

Figure 7.13: *Different representation of real world objects between datasets*

Haiti (Port-au-Prince area)	UN – GMM datasets		GMM-OSM datasets	
	Number	Percentage	Number	Percentage
Total outliers	35	8.56% of total tiles	14	3.58% of total tiles
Data matching errors	17	48.57	7	50.00
Distance errors	4	11.43	0	0.00
Different representation	14	40.00	7	50.00

Table 7.10: *Evaluation of positional accuracy outliers*

7.4.5. VGI commission indication

Sections 4.13.2 and 4.13.3 described how some non-matched VGI features are marked as possible new features, based on their attributes or road type. The tiles shown in Figure 7.11 were manually evaluated (see section 5.4.5 for more information on the evaluation) and results (Table 7.11) show error levels of 12.52% and 2.42% respectively for UN-GMM and GMM-OSM comparison. The first one is higher, reflecting the bigger inconsistencies between the datasets examined that lead to higher data matching errors, as also shown in Table 7.7. The lower data completeness of the reference dataset in both cases, on the other hand, leads to high percentages of correct indication of VGI commission (95.39% and 76.29%, in contrast to the two previous case studies, as described in Tables 5.10 and 6.21), which proves the efficiency of the method.

Evaluation of VGI Commission indication	UN-GMM: GMM dataset		GMM-OSM: OSM dataset	
	Number	Length (m)	Number	Length (m)
All non-matched objects	395	50,232,722 (100%)	2,422	174,103,480 (100%)
Indicated non-matched objects as new	389	47,918,169 (95.39%)	1,923	132,817,249 (76.29%)
Data matching errors (instead of new objects)	39	6,000,310 **(12.52%)**	35	3,209,673 **(2.42%)**

Table 7.11: *Evaluation of VGI commission indication*

7.5. Discussion

7.5.1. Data matching errors and quality results

Section 4.14 argued on the impact of data matching errors on quality results using the equations 11, 12 and 14. Similarly to the way they were applied on the two previous case studies (sections 5.5.1

and 6.5.1), Table 7.12 estimates how quality results of Tables 7.2, 7.3, 7.4 and 7.5 are affected by the data matching error levels of Table 7.7. Values in bold exceed the error ranges defined in section 5.5.1. They are attributed to the much greater dissimilarities in primary road names (attribute accuracy) and bigger distances for many corresponding objects (positional accuracy).

VGI spatial quality element	Estimated error range
Completeness (based on data matching)	<+0.91% (UN-GMM) , <+0.65% (GMM-OSM)
Attribute accuracy	<+ 0.34% (UN-GMM), >**- 7.62% (GMM-OSM)**
Positional accuracy	Outliers: <**9.38% (UN-GMM)**, <3.62% (GMM-OSM)

Table 7.12: *Estimation of errors in quality results for the provided method*

Leaving aside attributes, which are generally missing from one of the two datasets in each comparison case, the second case proves to provide more reliable quality results. However, although GMM was regarded as the reference dataset, mainly because its structure is closer to that of an official dataset, it is still crowd-sourced data. The quality results in the case of comparing two VGI datasets should be considered as an indication of their level of agreement, and not as an indication of completeness and accuracy, in terms of representing the real world objects.

7.5.2. Road types correspondence

UN matched to GMM			GMM matched to UN		
UN Road Type	**GMM Road type**	**%**	**GMM Road Type**	**UN Road type**	**%**
Primary	PRIMARY_HIGHWAY	36.1	LIMITED_ACCESS	Arterial Street	100
	MAJOR_ARTERIAL	19.3		-	-
Arterial Street	LOCAL	94.8	LOCAL	Arterial Street	84.2
	MINOR_ARTERIAL	3.7		Unclassified	12.1
Unclassified	LOCAL	95.6	MAJOR_ARTERIAL	Primary	43
	TERMINAL	2.4		Highway	38.4
Secondary Street	LOCAL	71.6	MINOR_ARTERIAL	Arterial Street	51.2
	MINOR_ARTERIAL	27.6		Highway	20.8
Highway	MINOR_ARTERIAL	36.9	NON_TRAFFIC	Arterial Street	99.7
	MAJOR_ARTERIAL	32.8		Unclassified	0.3
<no value>	LOCAL	100	PRIMARY_HIGHWAY	Primary	71.8
	-	-		Highway	22
			SECONDARY_ROAD	Primary	93
				Arterial Street	4.5
			TERMINAL	Arterial Street	55.7
				Unclassified	44.3

Table 7.13: *UN and GMM road types correspondence*

GMM matched to OSM			OSM matched to GMM		
UN Road Type	**GMM Road type**	**%**	**GMM Road Type**	**OSM Road type**	**%**
<no road type value>	footway	37.8	residential	Unclassified	99.2
	path	25		<no road type value>	0.2
Arterial Street	tertiary	46.9	track	Unclassified	96
	secondary	40.3		Terminal	2.4
Primary_Highway	primary	86.3	tertiary	Unclassified	92.9
	secondary	11		Arterial Street	6
Limited Access	service	62.8	path	Unclassified	92.7
	tertiary	37.2		<no road type value>	7.3
Primary	secondary	76.1	unclassified	Unclassified	97.9
	tertiary	12.6		Arterial Street	1.9
Secondary Street	secondary	66.9	service	Unclassified	88.5
	primary	30.2		Terminal	10.8
Terminal	track	48	secondary	Unclassified	32
	service	39.4		Arterial Street	28.4
Unclassified	tertiary	37.8	footway	Unclassified	73.3
	residential	31.8		<no road type value>	26.7
			primary	Primary_Highway	69.6
				Secondary Street	9.1
			road	Unclassified	100
				-	-
			pedestrian	<no road type value>	80.5
				Unclassified	19.5

Table 7.14: *GMM and OSM road types correspondence*

Similarly to section 5.5.2, road type correspondence information is collected (Tables 7.13 and 7.14). Corresponding road types are considered those which, when examined from each dataset point of view, they are matched primarily with the other road type. Thus it is a two-way link, bilaterally calculated. Road types without a two-way link have an ambiguous correspondence, which leaves the following general road type connections:

- UN 'Primary' with GMM 'Primary_Highway',
- UN 'Arterial Street' with GMM 'Local'.
- GMM 'Primary_Highway' with OSM 'primary'.
- GMM 'Unclassified' with OSM 'tertiary'.

7.5.3. VGI commissioned data

Information on what road types are generally mapped or failed to be mapped is presented in Tables 7.15 and 7.16. This information can be used to find commissioned data from both datasets. For the first case of comparison between UN and GMM datasets (Table 7.15), UN's 'Unclassified' road type refers to data that are mapped by GMM at less than 50%. On the other side, GMM 'Non-Trafic' road type generally refers to data not mapped by UN (at about 82%). For the second case (GMM and OSM, Table 7.16), the same GMM road type 'Non-traffic' is the less adequately covered by OSM, however to a much lower percentage (33%). OSM on the other hand has 33 road types, as a result of the decision not to follow any standards. The first 15 in terms of network length are presented, rejecting road types of length less than 500 m and unusual road type values (e.g. imp emmanuel, circonstancielle, Entree du plan). For both cases, however, the same conclusion can be made, also in accordance with the previous chapters: selected road types should not be rejected in order to create two datasets that will provide similar road type information, since it would lead to throwing away valuable data, which will further affect the final quality results.

UN – GMM datasets					
UN road type	**Matched length (m)**	**Non-matched length (m)**	**GMM road type**	**Matched length (m)**	**Non-matched length (m)**
Highway	56,907 (100%)	0 (0%)	PRIMARY_HIGHWAY	55,741 (99.7%)	168 (0.3%)
Primary	116,802 (97.12%)	3,460 (2.88%)	SECONDARY_ROAD	15,046 (73.85%)	5,328 (26.15%)
Secondary Street	50,248 (96.96%)	1,574 (3.04%)	MAJOR_ARTERIAL	51,682 (98.5%)	789 (1.5%)
Arterial Street	1549,380 (85.14%)	270,475 (14.86%)	MINOR_ARTERIAL	109,631 (97.78%)	2,493 (2.22%)
Unclassified	224,487 (48.42%)	239,126 (51.58%)	LOCAL	1654,022 (78.44%)	454,504 (21.56%)
<Null road type value>	274 (63.96%)	154 (36.04%)	TERMINAL	11,086 (62.57%)	6,632 (37.43%)
			NON_TRAFFIC	1,740 (18.15%)	7,848 (81.85%)
			LIMITED_ACCESS	158 (100%)	0 (0%)

Table 7.15: *UN and GMM road types: what is mapped by the other dataset and what is not*

Although it is suggested to use all road types during the data matching procedure, the above information on road type correspondence can be used in order to find VGI commissioned data. After

matching the necessary data regardless of their road type, the VGI non-matched dataset will probably include features of all road types. The ones with a high non-matched percentage refer to data not mapped by the reference dataset, so they can be ignored more safely. By this, non-matched VGI data volume that needs to be manually examined for commission is reduced. Alternatively, if such data are needed for conflation purposes (e.g. to add footpaths to a reference dataset), they can easily be isolated and used (e.g. Appendix C-Figure 12).

GMM – OSM datasets					
GMM road type	**Matched length (m)**	**Non-matched length (m)**	**OSM road type**	**Matched length (m)**	**Non-matched length (m)**
PRIMARY_HIGHWAY	56,448 (100%)	0 (0%)	residential	657,819 (70.08%)	280,841 (29.92%)
SECONDARY_ROAD	20,374 (100%)	0 (0%)	track	320,503 (34.89%)	598,182 (65.11%)
MAJOR_ARTERIAL	52,471 (100%)	0 (0%)	tertiary	847,491 (92.95%)	64,235 (7.05%)
MINOR_ARTERIAL	112,123 (100%)	0 (0%)	path	20,764 (7.4%)	259,883 (92.6%)
LOCAL	2,027,207 (95.76%)	89,837 (4.24%)	unclassified	172,645 (65.71%)	90,089 (34.29%)
TERMINAL	17,393 (98.17%)	325 (1.83%)	service	50,644 (26.27%)	142,148 (73.73%)
NON_TRAFFIC	6,461 (67.39%)	3,127 (32.61%)	secondary	169,642 (96.84%)	5,539 (3.16%)
LIMITED_ACCESS	158 (100%)	0 (0%)	footway	7,824 (6.49%)	112,677 (93.51%)
			primary	79,796 (97.88%)	1,726 (2.12%)
			road	595 (9.69%)	5,551 (90.31%)
			pedestrian	1,158 (4.61%)	3,546 (75.39%)
			raceway	4 (0.12%)	3,199 (99.88%)
			unspecified	388 (35.19%)	715 (64.81%)
			Residentielle	722 (76.27%)	225 (23.73%)
			residential; unclassified	858 (99.28%)	6 (0.72%)

Table 7.16: *GMM and OSM road types: what is mapped by the other dataset and what is not*

Section 5.5.3 noted that the manual examination can be easier when using satellite imagery as a background. Figures 7.14 to 7.16 provide examples of VGI commission for each pair of examined

datasets (7.14, 7.15 for UN-GMM and 7.16 for GMM-OSM), where yellow lines represent the reference and red ones the VGI non-matched dataset.

Figure 7.14: *UN-GMM datasets: Failed data matching due to distance. Mislocated reference dataset?*

Figure 7.14 shows that if Google Earth imagery is to be trusted, the UN dataset has some features mislocated (southern tiles), which result in failed data matching due to their long distance. While it can be argued that the satellite imagery is accurately placed when there is no relevant metadata, the fact that each of the involved VGI projects uses satellite images from a different provider (GMM from Google, OSM from Yahoo!), makes it less probable that the official data are correct and that both VGI datasets, along with the satellite images, contain errors partially and of such a magnitude. It can be argued, therefore, that the official dataset is not as immaculate as it should be. Appendix C – Figure 11 is a similar case. Figures 9, 10 and 12 of Appendix C are cases of GMM commissioned data.

When using Google Earth to check for VGI commission, as described in section 5.5.3, historical imagery can also be accessed, which is useful to check if commission also means more updated data. While KML files presenting the output datasets remain the same, the background satellite imagery can be rolled back before the earthquake. However, the above mentioned data shift errors remain more or less the same. Additionally, Figure 7.15 shows that U.N. data seem to have been collected around 2003, so they are far less updated than the two VGI datasets, which are mainly created just after the 2010 earthquake. This also explains the increased number of VGI commission cases.

Figure 7.15: *U.N. dataset with* ***a:*** *2003 Google maps image,* ***b:*** *2010 Google maps image,* ***c:*** *2010 Google maps image and GMM non-matched (commissioned) data (in red)*

Figure 7.16: *GMM-OSM datasets: OSM commission example*

For the second pair of data sources comparison (GMM and OSM), Figure 7.16 is an example of OSM superiority to GMM (VGI commission), while Figures 13 and 14 of Appendix C are similar cases.

Generally, in both the examined cases, it was relatively easy to find commissioned VGI objects, because the datasets used as reference (UN or GMM) were inferior to the VGI ones (GMM or OSM) in terms of data volume. However, manual examination is always necessary to exclude data matching errors, as well as data that should not be considered as commissioned due to their type and the reference dataset specifications.

7.5.4. Topology correction

Section 5.5.4 noted that low-quality topology in this context mainly refers to features that do not end at road junctions, while other cases include single features that may describe two different objects, or features with more complicated shapes than a crooked line (called 'multi-linestrings'). This leads to inconsistent corresponding objects that should only be partially matched, however when examined as entities it is difficult to decide for the feature correspondence even in cases of manual data matching.

Some examples are presented in Figures 7.17a to 7.17c, where labels refer to the feature ID and are used to show where each feature starts and ends: In Figure 7.17a UN feature 4640 (inside grey ellipse) should only partially be matched with GMM 6232. UN feature 1165 fails to be partially matched with GMM feature 6235 (inside yellow ellipses) due to its size. The same applies to GMM

dataset in the southern part, where feature 2500 fails to be matched with the UN's 6235 (inside blue ellipse). Figure 7.17b shows how the above can lead to different feature shapes that affect feature correspondence. Two features are selected from each dataset and buffered. To the north, GMM feature 3091 could link to UN's 229, 228 or part of 220. To the south, UN's 376 and GMM's 3090 end-points differ significantly. In the centre, UN's feature 220 is also mistakenly matched where there is no corresponding GMM object. Finally, Figure 7.17c shows a selected multi-linestring feature in UN (reference) dataset (the blue one). As section 7.2 noted, such topological errors in datasets created the need for some data preparation, according to which all intersected features were divided in order to enable a more efficient feature matching.

Figure 7.17: *Topological errors,* ***a and b:*** *Features not divided at road intersections,* ***c:*** *Use of multi-linestring*

To justify the above data correction, the method was also applied on the original UN and GMM datasets, repeating the evaluation without any topological correction. The same tiles of Figure 7.11 were again manually evaluated. The different error levels between topologically corrected datasets (Figure 7.12a) and non-corrected datasets (Figure 7.18) highlight the importance of the data preparation followed in section 7.2. Compared to Table 7.7, Table 7.17 proves that the method is less efficient when topological errors similar to the described ones occur and are not considered.

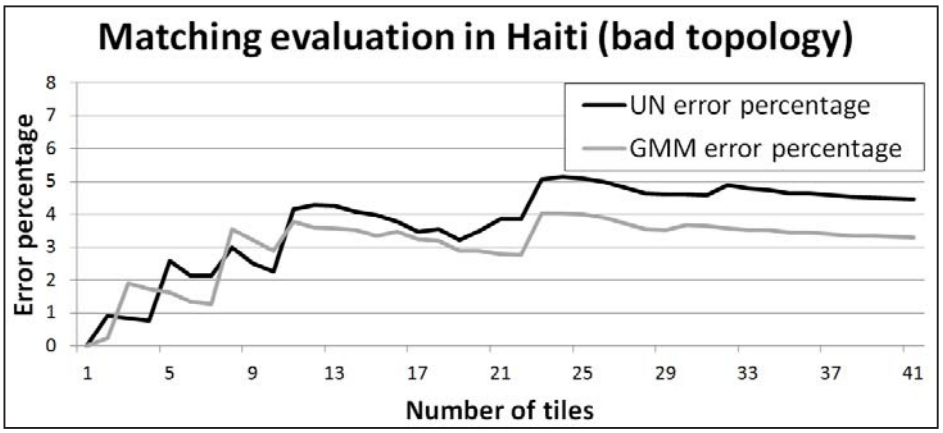

Figure 7.18: *Data matching error levels between topologically corrected datasets and original ones*

Data-set	*Area*	*Length (m) compared*	*Length (m) evaluated*	*Missing length (m)*	*Surplus length (m)*	*Total matching error (m)*
UN	Haiti	2,490,840	298,319 (11.8%)	13,291 (4.46%)	0 (0%)	**13,291 (4.46%)**
GMM		2,372,981	289,207 (12.2%)	7,959 (2.75%)	1,571 (0.54%)	**9,530 (3.30%)**

Table 7.17: *Data matching errors between UN and GMM datasets topologically non-corrected*

7.6. Summary

In this chapter the proposed framework proves its generality by being applied in different data sources and area, created for different purposes than in previous chapters. Results show that the method works efficiently, with error levels more or less similar to the ones of previous chapters.

The existence of official data and VGI from two crowd-sourced providers for the same area further enables the evaluation of the two VGI sources, with one of them considered as reference dataset. Results prove that the method also works efficiently, which adds a new dimension to this research:

the proposed framework can also be used to compare VGI sources when official data are not available.

Compared to the previous case studies where both datasets had a high-quality topology, this chapter raised questions on topology correction importance, as it proved that incorrect topology leads to erroneous or even ambiguous data matching: there are cases where it is difficult to decide whether to consider two features as corresponding ones when they are only partially matched. Correcting the topology by splitting features at their intersections reduces data matching errors significantly, however it does not solve the problem of different feature representation in all cases.

Attribute accuracy evaluation proved to be less efficient due to the nature of the data. Road names differ significantly between the datasets, which leads to rejecting similar strings due to their low level of similarity.

The similar scope and objectives between the reference and VGI datasets used in this chapter, along with the reduced coverage of the reference dataset in each case, leads to a more fruitful discovery of commissioned VGI data. However, they still need to be manually examined, in order to reject objects that were mistakenly not matched or they are not designed to be represented by the reference dataset, in which case they should not be considered as commissioned data. In cases where this specific data type is sought as well, e.g. for conflation purposes, the cumbersome manual examination can be reduced.

Combining the results of both case studies along with the use of satellite imagery as background, the two VGI datasets are in accordance with each other and with the satellite imagery for all data, while the official dataset from UN seems to have some features mislocated by more than 100 m. It would be far too daring to suggest that this method could be also used to test the validity of official data, however if there are different VGI sources that agree to each other but not with the official dataset, this surely raises some interesting questions. Additionally, it was found that data collected for VGI datasets are more recent than the ones used in U.N. dataset, so VGI sources prove to be more up-to-date.

Chapter 8

Discussion

8. Discussion[7]

8.1. Introduction

The three case studies that were presented in previous chapters suggest that the proposed methodology is generally applicable, robust and efficient. Each case included a relevant discussion, triggered by each analysis. This chapter moves away from the individual case study level to a broader view by further discussing and exploring some more general aspects of the methodology. Additionally, this broader view contributes in understanding the implications of quality on VGI, as well as in defining the limitations of the methodology. Successively, potential usage of this framework is discussed, linking back to the motivation described in the beginning of the thesis (section 1.8).

8.2. VGI heterogeneity and tiling approach

Section 4.6 mentioned that VGI heterogeneity can be dealt by splitting data into tiles in order to examine them separately and produce individual results, while section 4.15.1 discussed the tile size and shape. Section 4.6 further described the use of extended tiles to address the issue of data so close to the tile boundary that their corresponding ones of the other dataset lie in the adjacent tile. Finally, section 4.7 described how tiles are classified differently according to their network density, in order to differentiate the data matching constraints between rural and urban areas. There are, however, three questions regarding the efficiency of dealing with heterogeneity and producing robust results:

1. Are results independent of the tile size and place? In other words, by splitting data differently, would the method produce the same results?
2. Is anything gained by extending the tiles, which increases the computational time by processing data that eventually will be discarded? Are tiles appropriately extended, succeeding in matching data next to their borders while at the same time keeping the additional computation at low levels?
3. Are tiles properly classified as ‘urban’ or ‘rural’, so that the use of looser data matching constraints is more successful in finding corresponding objects in larger distances due to the lower positional accuracy in rural areas?

[7] Section 8.3 has been partially adapted from:
Koukoletsos, T., Haklay, M. and Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, [in press - DOI: 10.1111/j.1467-9671.2012.01304.x].

8.2.1. Tile selection

To answer the first question, the same VGI and reference data were processed for the first case study, using a different grid size, as well as the same size of grid slightly shifted. Specifically, a new grid of 4 km^2 was created (2 km x 2 km), and additionally the initial 1 km^2 grid was slightly and randomly shifted (295.25 m in x and -348.77 m in y direction). The areas of the first case study (Greater London and west of Newcastle) were selected because they are relatively small and they represent two different types of network (urban or dense and rural or scarce), which will show if there is a different behaviour in each case.

There are however some difficulties in finding what to compare. In each case, each tile will have different matching results due to the different data involved, which will affect data completeness, positional and attribute accuracy. As an example, Table 8.1 provides the differences between quality results for 35 tiles randomly selected in the centre of London: while average differences in lengths, percentages or positional accuracy for these tiles are quite low, individually they may vary significantly. Additionally, tiles with no data from both datasets are neglected, as comparison is not possible. If, however, a different tile is used and data now exist from both datasets, the previously omitted tiles and data will now be added on the comparison procedure. Obviously no corresponding features will be found for these objects, as the other dataset contains no data in this area, however the compared lengths (explained in sections 5.2 and 5.3) will change. Figure 8.1 shows an example; when 1 km^2 tiles are used, there is nothing to compare for cells NY7288, NY7388 and NY7387, so these tiles are neglected. In case of using 4 km^2 tiles, the same data, included now in cell 5343111, will be examined in the matching process. So, there seems to be no useful indicator in individual tile results or length.

What, however, should remain the same regardless of the tessellation method, is the total matched length. Expressed as percentage, the total matched length should be compared with the initial dataset's length (and not with the sum of the lengths of the processed tiles, called as 'length compared'). Tables 8.2 and 8.3 for rural and urban areas respectively show the total compared and matched lengths for the three tessellation scenarios. The independency between the proposed matching approach and any data splitting method is proven by the similar total matching results, despite that the total compared length may differ (especially in rural areas); differences in matching are less than 0.41% for rural and 0.11% for urban areas.

Tile ID	ITN length (m)	OSM length (m)	ITN match percentage	ITN name1 percentage	ITN name2 percentage	ITN names percentage	OSM match percentage	OSM name1 percentage	OSM name2 percentage	OSM names percentage	OSM pos. accuracy (m)
TQ2578	3.4	-2.7	-0.62	0.28	-0.02	-0.17	20.3	0.14	0	0.17	-0.13
TQ2579	-2.8	0	0.91	-0.38	-0.4	-0.38	-7.62	-2.72	0	-2.52	-0.44
TQ2580	8.4	1.2	0.46	0.4	-7.34	-1.06	31.17	5.87	0	4.24	-2.56
TQ2581	-3.5	0.4	0.05	-0.16	8.93	1.37	-14.7	-0.1	3.69	0.71	-0.38
TQ2582	-1.6	-3.4	-0.31	-0.14	-16	-2.27	7.26	-0.58	0	-0.54	0.5
TQ2583	0.8	1.7	0.75	-0.63	-10.2	-2.83	-0.63	-3.01	-10.5	-4.67	1.125
TQ2584	-0.6	0.9	4.86	5.85	32.57	10.15	-3.62	3.06	32.9	7.66	-0.63
TQ2678	1.5	0.5	0.1	-0.45	9.65	0.83	2.66	0.1	0	0.16	0.844
TQ2679	-7	0.5	-0.53	0.07	25.44	3.78	-27.2	-11.3	0	-10.2	0.438
TQ2680	4.9	-1.7	-2.67	-4.9	-29.4	-10.1	21.38	43.46	-17.1	37.11	-8.25
TQ2681	1.6	2.8	1.14	-0.29	3.77	-0.1	-1.87	0.24	-0.21	-0.07	-1.13
TQ2682	-6.4	-10.2	0.91	4.79	24.04	9.84	9.51	1.41	15.84	4.7	-0.38
TQ2683	0.2	0.8	-2.81	-2.38	-3.75	-2.63	-3.31	-1.85	-4.2	-2.28	1.5
TQ2684	-0.3	-1.9	0.02	-0.1	1.5	0.16	3	3.18	-0.94	2.57	0.25
TQ2778	4.5	4.6	-0.26	-1.08	-11.1	-1.85	-3.29	-0.36	-2.27	-0.58	-2.63
TQ2779	-5.6	2.6	0.01	0.86	3.39	0.69	-31.5	-10.3	5.22	-8.19	0.313
TQ2780	-1.5	-0.8	-1.16	2.25	-22.6	-3.54	-1.53	10.99	-3.84	7.62	-0.25
TQ2781	6.1	-0.1	-0.78	-1.63	12.91	0.61	16.17	2.06	-2.93	1.1	-0.25
TQ2782	-5.9	-6.4	-0.39	-0.92	19.06	3.3	-5.51	-4.82	0	-4.12	0.125
TQ2783	4.9	2.2	3.21	-0.93	-7.11	-1.89	21.29	10.61	0	8.63	-0.94
TQ2784	2.9	-1.1	-0.14	0.52	-6.52	0.91	18.41	2.56	-5.37	2.42	-0.56
TQ2878	3.6	2.1	0.13	-0.66	5.53	0.43	4.7	-0.38	-5.58	-1.25	0.5
TQ2879	-0.4	5.2	0.19	-3.08	-9.76	-5.66	-14.2	-2.49	0.1	-1.78	-0.38
TQ2880	2.9	-1.5	0.2	2.03	-3.77	3.67	11.03	-1.75	1.17	-2.04	-0.56
TQ2881	0	-1.1	-0.42	-0.26	9.5	0.83	2.72	1.3	0	1.33	-0.63
TQ2882	-9.3	-2.3	-0.01	0.37	19.04	4.13	-28.4	-2.95	4.61	-1.79	-0.38
TQ2883	-2.9	-2.5	0.16	0.41	-8.41	-1.07	-4.7	-5.82	0	-5.12	0.125
TQ2884	-0.6	-0.4	-0.18	-0.36	4.19	1.11	-0.07	3.26	-5.84	1.26	0.5
TQ2978	8.5	5.9	-0.39	0.39	8.33	3.27	12.26	4.51	17.93	8.4	0.031
TQ2979	-4.8	2.8	1.23	0.33	3.79	1.49	-29.5	-3.38	-1.2	-3.22	0.75
TQ2980	6.2	3	1.2	0.49	21.99	10.25	13.96	-0.69	0	-0.51	-2.88
TQ2981	-1.8	-3.2	-0.12	0.63	2.58	1.37	3.19	5.11	0	4.65	0
TQ2982	0	-1	-0.55	-0.33	3.69	0.59	3.06	1.34	-5.67	-0.13	0.75
TQ2983	1.1	0.1	2.52	1.38	8.05	3.19	7.44	2.88	0	2.5	-0.75
TQ2984	4.9	4.7	0.34	0.09	0.76	0.21	4.71	2.8	0	2.42	0.5
Average	**0.3**	**0.0**	**0.06**	**0.05**	**0.78**	**0.19**	**0.60**	**0.84**	**0.38**	**0.78**	**-0.29**

Table 8.1: *Differences for 35 tiles in central London because of tile shifting: quality results minus quality results for shifted tiles*

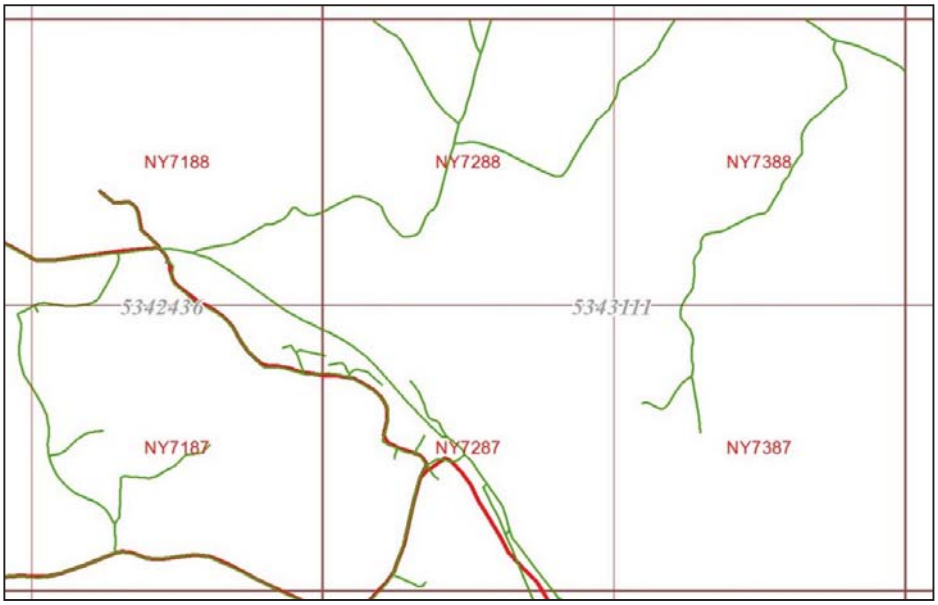

Figure 8.1: *Variations in data comparison due to data splitting*

Datasets & total lengths (m)	Lengths	Tile size 1 km^2	Tile size 1 km^2 - slightly shifted	Tile size 4 km^2
OSM (rural area): 1,872,836	Compared	1,847,186 (98.63%)	1,854,616 (99.03%)	1,871,426 (99.92%)
	Matched	**1,678,990 (89.65%)**	**1,671,248 (89.24%)**	**1,672,657 (89.31%)**
ITN (rural area): 2,935,672	Compared	2,459,594 (83.78%)	2,466,215 (84.01%)	2,655,899 (90.47%)
	Matched	**1,717,604 (58.51%)**	**1,712,431 (58.33%)**	**1,707,557 (58.17%)**

Table 8.2: *Compared and matched lengths in rural area for different splitting methods*

Datasets & total lengths (m)	Lengths	Tile size 1 km^2	Tile size 1 km^2 - slightly shifted	Tile size 4 km^2
OSM (urban area): 20,229,391	Compared	20,214,515 (99.93%)	20,214,216 (99.92%)	20,227,239 (99.99%)
	Matched	**16,568,469 (81.90%)**	**16,567,294 (81.90%)**	**16,545,230 (81.79%)**
ITN (urban area): 18,368,148	Compared	18,366,914 (99.99%)	18,366,895 (99.99%)	18,368,148 (100%)
	Matched	**16,983,470 (92.46%)**	**16,970,023 (92.39%)**	**16,970,287 (92.39%)**

Table 8.3: *Compared and matched lengths in urban area for different splitting methods*

8.2.2. The use of extended tiles

To answer the second question of section 8.2, the selection of the 50 m buffer that extends each tile side by 100 m (according to which data are clipped) is a result of the following considerations:

- The maximum search distances used in data matching are 26 and 51 meters respectively for urban and rural areas (section 4.8.3). Above that, objects are not considered as

corresponding ones, so the use of a larger buffer would include additional data that would not be examined for correspondence.

- The directional analysis that is performed in the data matching stage is applied on the segments and depends on the length of the segment, as described in section 4.8.3. When a segment close to the tile border is clipped, the angular tolerance for finding a corresponding object with similar orientation increases according to the graphs of Figure 4.4. This may lead to erroneous data matching. By using the extended tile, segments vertical and close to the tile border which are shorter than 50 m will not be clipped at all, so directional analysis is not affected. In cases of non-vertical segments, this approach can also handle longer segments. For the worst case scenario of vertical segments longer than 50 m, however, Figure 4.4 shows that angular tolerance does not change significantly, so this threshold is sufficient to produce the same results regardless of the tiling method. The additional objects between the initial tile border and the extended one (or even beyond it) that will also be directionally examined may be mismatched, because actually the problem moves to the new extended border. However, they will all be rejected when clipping the examined data to the initial tile border.
- The length of segments of the datasets involved in the first case study was selected to be examined, because it refers to dense and scarce networks individually, so the collected information is more representative of the differences between them. Figure 8.2 presents the length frequency for the datasets and areas examined, while Table 8.4 shows the relevant descriptive statistics.

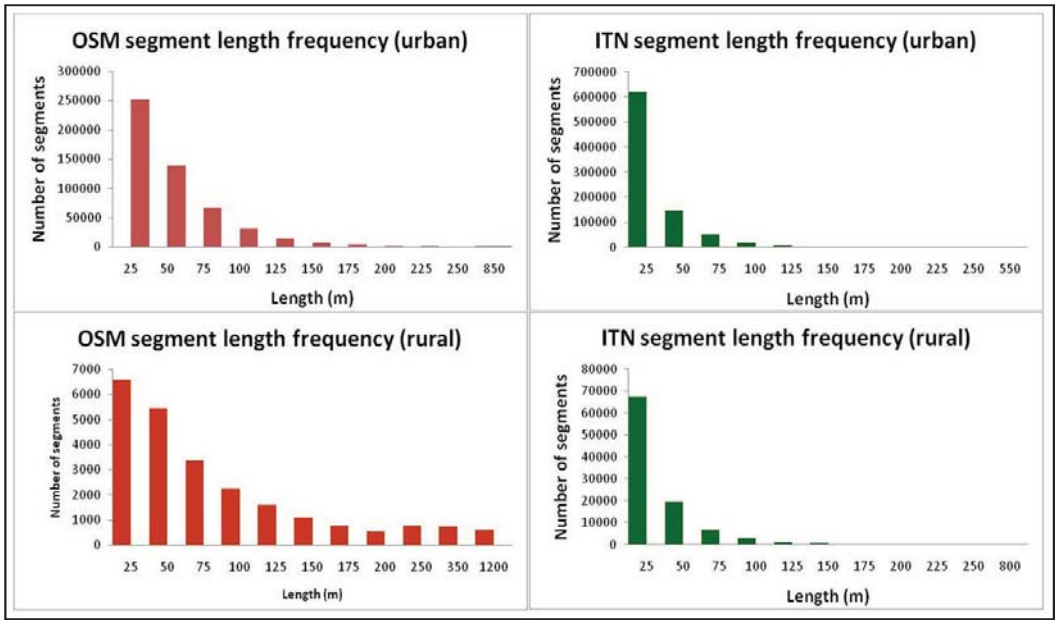

Figure 8.2: *VGI and reference segment length frequency of first case study*

	Urban area (Greater London)		Rural area (West of Newcastle)	
	ITN (reference)	OSM (VGI)	ITN (reference)	OSM (VGI)
Max	528.867	838.350	764.258	1195.54
Average	21.357	39.162	25.208	81.217
Median	12.083	26.518	15.222	48.972

Table 8.4: *VGI and reference segment length statistics (m) for the first case study*

- The extended area needs to be of such a size that will not delay the whole process significantly, as the same data will be examined many times. With a tile size of 1 km^2, a buffer of 50 m leads to 21% larger tile (each side is 100 m larger), which computationally is reasonable, compared to using a higher buffer value.

The use of the extended tile improves data matching, which successively leads to a better assessment of data completeness, attribute and positional accuracy. An example is provided in Figure 8.3. VGI feature that runs parallel to the south border of tile NY7164 is found with no corresponding object when the initial tile size is used (left figure), because its corresponding reference object, although very close to it, lies in the adjacent tile. When using the extended tile (right figure), data matching is improved. Table 8.5 describes the improved data completeness and attribute accuracy. Corresponding results for the reference dataset remain the same and are not presented. In Figure 8.3, however, positional accuracy is not significantly affected, but it would have been much different if the VGI and reference features had the opposite places. Figure 8.4 provides such an example.

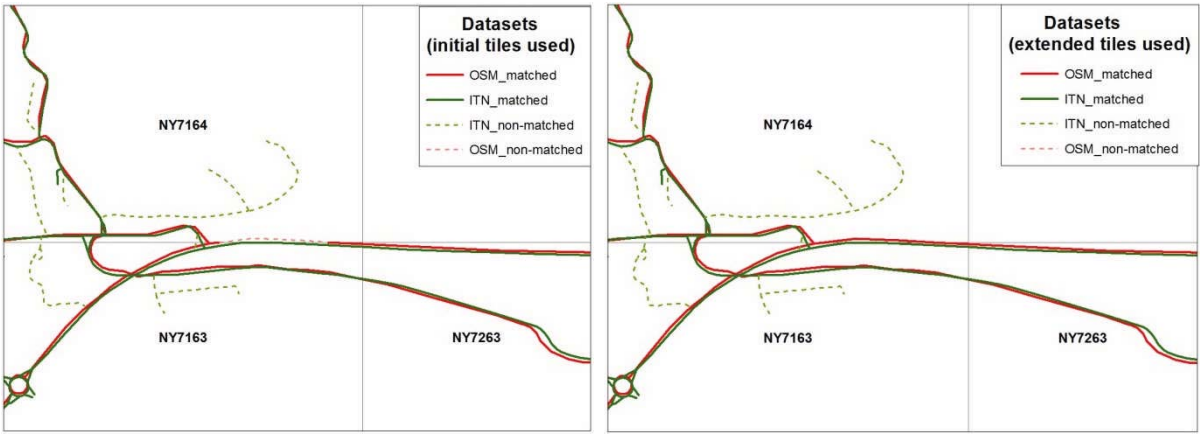

Figure 8.3: *Benefits in data matching when using the extended tile*

Tile 'NY7164'	Use of initial tile size	Use of extended tile size
VGI matched %	90.93%	**99.92%**
VGI Name1 accuracy	71.35%	71.35%
VGI Name2 accuracy	67.97%	**100%**
VGI total attribute accuracy	69.22%	**89.39%**
VGI Positional accuracy	7.6875 m (95%)	7.875 m (95%)

Table 8.5: *Benefits in quality evaluation when using the extended tile*

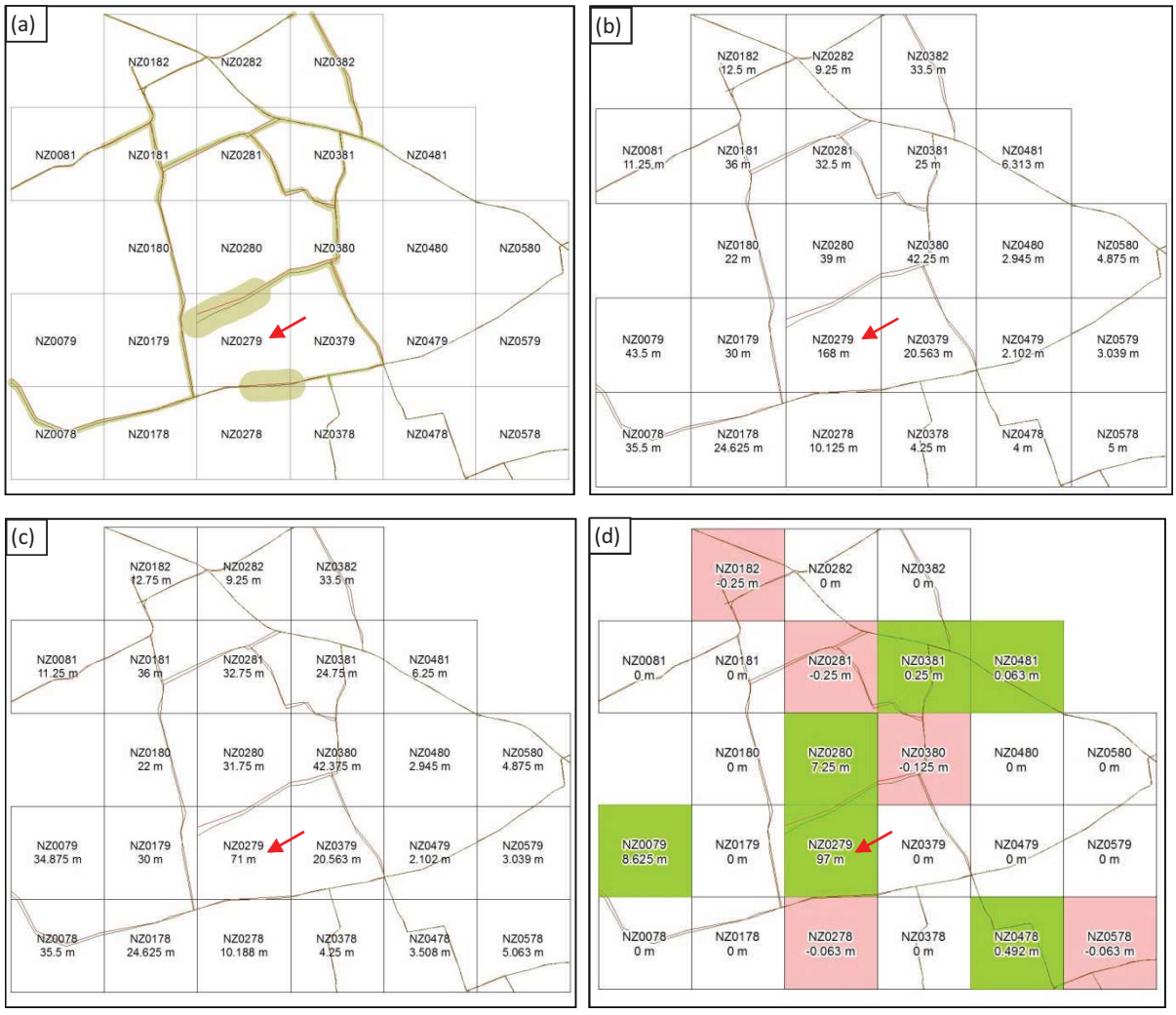

Figure 8.4: *Benefits in positional accuracy when using the extended tile*

Figure 8.4a shows the matched datasets and the buffer widths when no extended tile is used. In Figure 8.4b the buffer values show that tile NZ0279 has an abnormally low positional accuracy (168 m), which is not representative of the included data. This is because the reference road that is parallel to the south tile border crosses to the adjacent tile, so the VGI corresponding object is

partially alone in the tile (see Figure 4.11 for a closer view). Figure 8.4c shows the buffer values when the extended tile is used, where the mentioned tile has a more reasonable value (71 m), considering the lower accuracy of the other object to the north of the tile. Figure 8.4d shows the buffer differences between using the extended tile and not using it at all. Positive differences (green cells) reach 97 m (for the above mentioned tile), while there are also other tiles that benefit from the extended tile. The reason for the minor negative differences (0.25 m or less – pink cells) is the tolerance for the desired percentage (discussed in section 4.12.2), which is reached differently and the buffering procedure stops earlier.

Finally, manual evaluation of attribute accuracy in all case studies found no error due to the tiling method, which means that the 50 m border extension is sufficient for the attribute accuracy assessment.

8.2.3. Tile classification

To answer the third question of section 8.2, a trial-and-error empirical approach was used to deal with the necessity of applying different constraints in urban or rural areas due to the lack of relevant metadata (discussed in section 4.7). The rural dataset from the first case study was examined tile by tile for the total length, number of features (linear data) and junctions (point data) included, considering only tiles with data from both datasets. The area type it refers to is mainly rural, however it also includes some small-sized built-up areas with a network of increased density that could be described as urban (Figure 5.6).

Generally, a relatively scarce network tends to have less features and junctions and total length per tile. To decide on the threshold, the rural area was processed twice, firstly using the stricter constraints described in the data matching approach, and secondly using the looser ones (described in section 4.8). The latter led to an increased matching percentage, however it was not always the result of correct data matching. This is because not all data in rural areas have a lower positional accuracy that increases the distance with their corresponding ones, so in cases of better accuracy, non-corresponding data were mistakenly matched.

For reasons of simplicity, the extended tiles were not used. There are two cases that were examined manually; tiles that the looser constraints proved helpful (called ‘fixed’), and those that added wrong matching information (called ‘new errors’). For the first case 109 cells were selected, while for the second one 22 (Figure 8.5). For these cells and for both datasets the total length, number of features

and number of junctions were calculated, in an effort to find a pattern that will help classify a tile as rural or urban. Table 8.6 shows the necessary results and statistics.

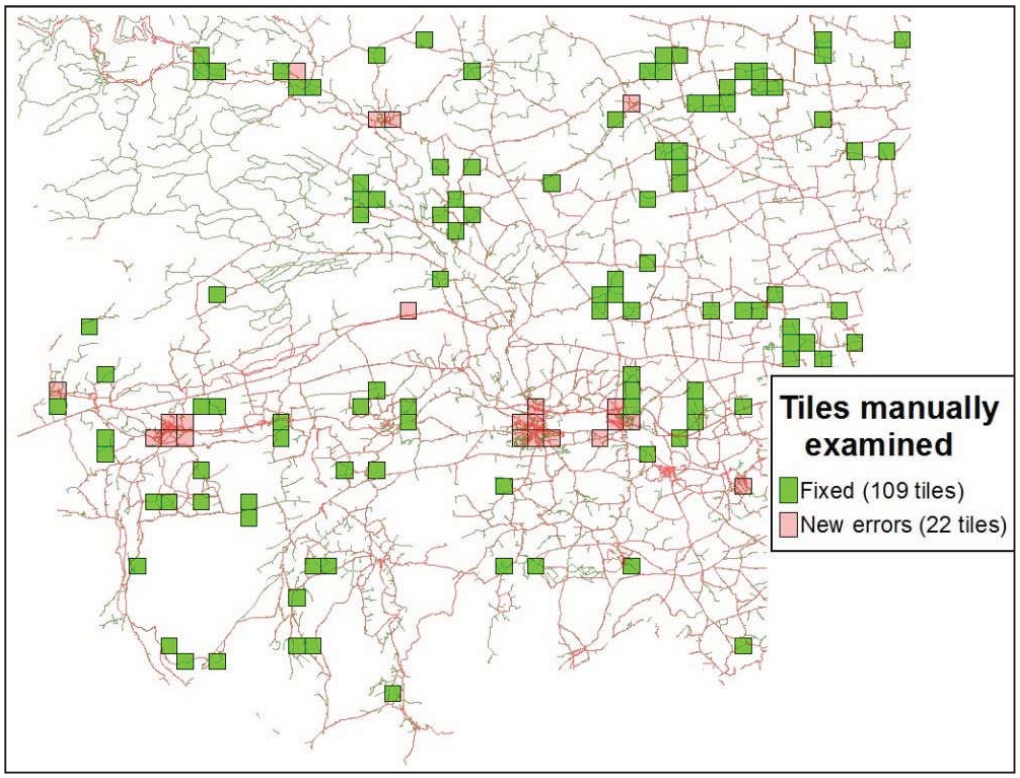

Figure 8.5: *Manually examined cells where by applying rural only constraints, data matching was fixed or deteriorated*

Statistics	Dataset	Length / cell	Features / cell	Junctions / cell	Dataset	Length / cell	Features / cell	Junctions / cell
Maximum	ITN fixed (109 cells)	7,597.137	42	17	OSM fixed (109 cells)	4,734.961	16	12
Minimum		164.879	1	0		53.863	1	0
Average		2,073.170	**8.596**	**3.569**		1,675.273	**3.028**	**1.202**
St.Deviation		1,111.914	8.145	4.077		915.014	2.425	1.704
Maximum	ITN new errors (22 cells)	13,647.997	227	121	OSM new errors (22 cells)	9,919.564	71	73
Minimum		1,014.125	1	0		686.399	1	0
Average		6,843.277	**85.500**	**43.500**		4,554.470	**21.682**	**18.500**
St.Deviation		3,291.305	62.575	34.576		2612.359	22.542	23.549
Maximum	ITN rural area (all cells)	13,647.997	227	121	OSM rural area (all cells)	9,919.564	71	73
Minimum		1.019	1	0		5.633	1	0
Average		1,711.762	**7.338**	**3.034**		1,446.206	**2.916**	**1.323**
St.Deviation		1,221.134	13.609	7.256		1,024.136	5.011	4.707

Table 8.6: *Statistics on total length and number of features and junctions when rural only constraints are applied in the rural area of first case study*

What is obvious is that additional matching errors ('new errors' in the table) usually refer to cells with much more features and junctions, as well as total length from the average of total or 'fixed' cells. So, the threshold that would divide rural from urban areas was decided to be based on the number of features and junctions.

The threshold is defined by the results of the reference dataset, as it is expected to be more complete, accurate and topologically correct. Rural areas was decided to be defined as the ones that have less than 17 features and 8 junctions per km^2. These numbers are the rounded results of adding the standard deviation to the average number of features and junctions respectively for the 'ITN fixed' category. Such a threshold will include 84% of the tested 'ITN fixed' population, as well as less than 15% of the 'new error' case, so it is expected to solve many cases of corresponding objects in greater distance with relatively small errors in missmatching. The same criteria are also applied on the OSM dataset, which leads to classifying tiles as rural ones where both reference and VGI datasets contain less than 17 features and 8 junctions per km^2.

The above threshold was used for the urban and rural areas of the first case study and classified tiles equally, as shown in Table 8.7; urban and rural areas include approximately 90% of urban and rural tiles respectively. Table 8.8 shows the differences in matching percentages when using urban only constraints (the stricter ones), compared to the mixed type that uses the tile classification. The efficiency of the selected threshold is proven by the total matching percentages in the tested rural and urban areas; final matching results are more or less the same for urban areas, while for rural areas matched features increase more significantly.

Area	Total tiles compared	Urban tiles	Rural tiles
Rural	1,299	138 (10.6%)	1,161 (89.4%)
Urban	1,687	1515 (89.8%)	172 (10.20%)

Table 8.7: *Tile classification of the first case study*

Dataset	Area	Matching % for urban only constraints	Matching % for mixed constraints
OSM (VGI)	Rural	84.5	88.1
	Urban	80.6	80.6
ITN (reference)	Rural	52.0	59.1
	Urban	92.2	93.0

Table 8.8: *Matching percentages with and without tile classification*

The above thresholds that distinguish rural and urban tiles are based on ITN and OSM datasets. Although they proved suitable also for the different datasets of the Haiti case study, further research is required to develop automatic recognition in the general case. Additionally, the selected thresholds and looser constraints are rather conservative, favouring the urban areas to avoid using looser constraints when this may lead to data matching errors. The manual evaluation of object matching efficiency in all case studies showed that there still are corresponding objects at even greater distances that are not matched (e.g. Figures 5.20 and 7.14). However, from a different point of view, the proposed method regards such objects as new ones (commissioned data), which after a closer manual examination gives an indication of the error in the datasets used. Such inconsistencies or errors need to be visible and to be considered accordingly, depending on the user requirements.

The examination of sections 8.2.1, 8.2.2 and 8.2.3 suggests that the method is robust and efficient regardless of the tile size or position. Tiles seem to be appropriately extended to negate the problems related to data splitting, and they are also quite appropriately classified for the use of different data matching constraints. Hence, the final results are more accurate and they succeed in providing a better insight of VGI heterogeneity.

8.3. Data matching and stages order

The order of the data matching stages (section 4.8) is set to deal with the nature of the data, each one trying to reduce the number of objects that need to be processed by the next stage. Stage 1 matches the obvious and leaves the ambiguous cases for next levels, based on simple geometric rules. Road names are not examined even if they are available, as '1-1' matching means that there is no other possible candidate to be matched, hence it is considered enough at this stage. The next stage sorts out correspondence by looking at exact name matching along with geometry, and is quite fast due to its reduced complexity. In urban areas where road network is denser, this stage succeeds in matching the biggest proportion of data, taking advantage of the fact that VGI is usually more complete in terms of road name attributes. By looking for text similarity, stage 3 is of an increased complexity, however previous stages should have significantly reduced the size of data to be examined. Stage 4 is also complex, as it is based on geometric calculations. In urban areas, data remaining to be examined are already minimised. In rural areas, however, VGI usually lacks road name information, so stages 2 and 3 will not be as successful as this one. The reduced network density in rural areas, on the other hand, keeps the number of objects for stage 4 relatively low. Stage 5 moves back to the feature level, linking the segment-by-segment examination of the reference dataset to the feature examination of the VGI dataset to follow. The next two stages (6

and 7) switch to the examination of VGI against reference features, instead of the opposite direction of the first four stages, providing a more thorough object examination between the datasets. Stage 6 comes before stage 7, as it is computationally less complex.

Case studies 1 and 3 show that the above stage order succeeds in matching data effectively, shifting the computational burden and stage importance accordingly (Tables 5.6 and 7.6). The quick (in computational terms) stage 2 does most of the work in the dense network of Greater London, where attributes are present. When attributes are sparser, the importance of stage 4 increases significantly for the rural UK area. In the Haiti case, finally, lack of attributes shifts the burden primarily to stage 4, which is computationally more demanding. The lack of attributes leads to a poorer geometric data matching (compared to when combined with the attribute constraints), which seems to be compensated by the contribution of the final stages: their contribution is much more significant than in the UK cases, where they rather correct minor errors (see section 5.4.1).

The efficient sequence and necessity of the stages is also implied by the data matching error levels found in all case studies (Tables 5.6, 6.17 and 7.6), which are relatively low and of similar scale size, regardless of the differences in stages contribution of each case study.

8.4. Selecting target percentage for positional accuracy

Section 4.12 mentioned that the user-defined desired overlap percentage is set to 95%, used as a value in all case studies, and that this can be regarded as a level of confidence for positional accuracy, because it describes the VGI data percentage that lies within a specific distance from the reference dataset. However, is this percentage suitable to lead to a representative buffer value? Why not testing the whole VGI source by using 100%?

Using 100% as target percentage is not suggested mainly for two reasons:

- Real objects are captured differently between the two datasets. Corresponding features may have different lengths (Figures 5.23b and 7.13). The purpose of the increasing buffer is to deal with the distance between two corresponding objects and not with their differences in length; when using 100 as a target percentage, buffer width is abnormally increased to include the full length of the VGI feature, which actually is not an issue of positional accuracy.
- Data matching may not be 100% correct for a tile. When a VGI feature is incorrectly classified as matched, the buffer that is applied to the closest (but not corresponding)

reference feature will have to be extended so that it will include all the VGI features, which leads to an abnormally high buffer value for the tile, non-representative of its positional accuracy. If, for example, an average matching error is 5%, deciding on an overlap percentage above 95% will obviously include part of the matching errors and positional accuracy estimation will be worse (bigger buffer value). Although this does not imply a higher risk for the decision-maker, it is not a representative value and the buffer indicator becomes very sensitive to outliers.

In order to decide on a suitable target percentage, two areas of 25 km^2 were selected with dense and scarce road network correspondingly (in Central London and west of Newcastle, Figure 8.6). Positional accuracy was applied for 11 different target percentages, starting from 90 to 100% using a 1% step. The extended tile was not used for reasons of simplicity and to also examine the extended tile contribution (already discussed in section 8.2.2). What is sought is a maximum percentage, above which buffer values increase abnormally to include matching errors or different feature lengths.

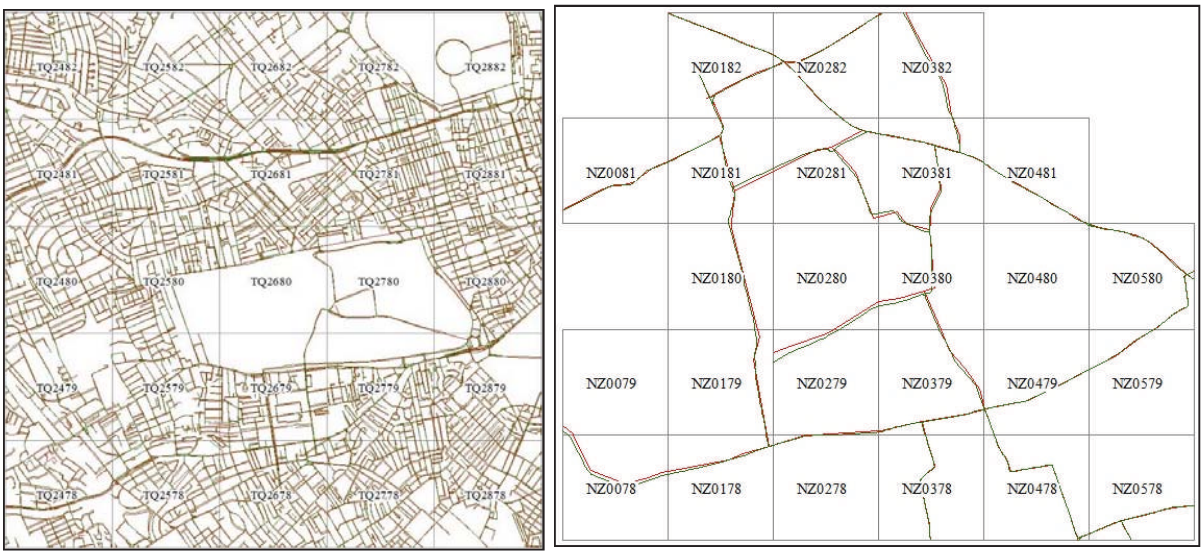

Figure 8.6: *Urban (left) and rural (right) areas tested to decide on target percentage*

Figure 8.7 shows how buffers increase while percentages move towards 100. In rural areas one tile had to be rejected as outlier due to abnormally high buffer values (tile NZ0279, buffers starting from 108 m, explained in section 8.2.2). A sudden change in buffer width can be easily spotted in urban areas at 98%. In rural areas such a change is difficult to spot due to far bigger buffer values, as a result of reduced positional accuracy.

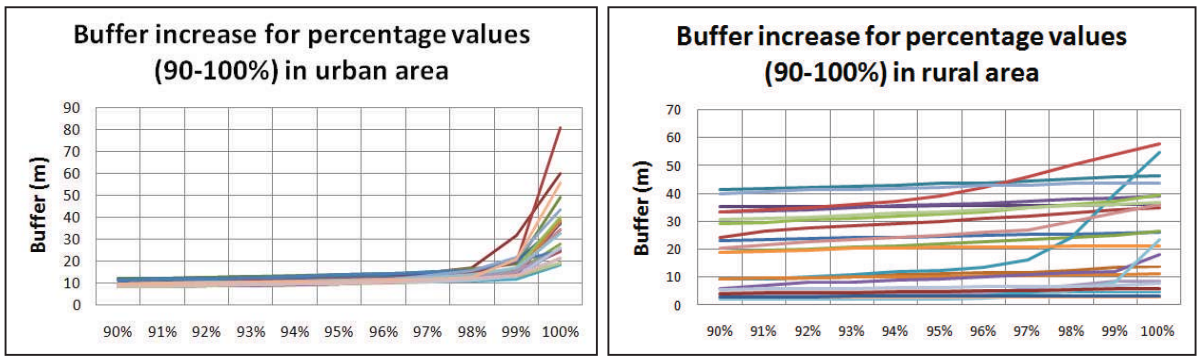

Figure 8.7: *Buffer widths corresponding to 1% increment of target percentage (90-100%)*

By calculating the differences between successive buffers, a median and average value of buffer increment was computed. This increment shows how positional accuracy (buffer width) is affected when the confidence level (user-defined desired percentage) is raised by 1%. As shown in Figure 8.8, average values are higher for rural areas due to the lower accuracy of some tiles. Median value provides estimation less susceptible to outliers, however results for the rural area are still far less normalised than the urban ones. Figure 8.8 skips percentage increments above 98%, as their magnitude would suppress and flatten the presented graphs.

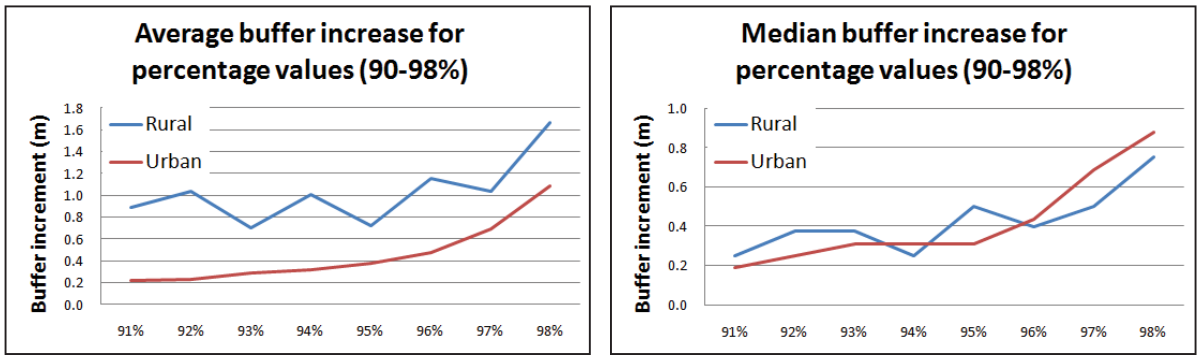

Figure 8.8: *Average and median buffer increment corresponding to 1% increment of target percentage*

A value of 95% is considered as the highest reasonable target percentage, since above that the buffer width may increase more than 0.5 and 1 m for urban and rural areas accordingly. It is also reasonable to cover data matching errors, which are found less than 5% for all case studies. Additionally, since the user-defined target percentage can be considered as a level of confidence for positional accuracy assessment, 95% is quite a high value that sounds satisfying for decision-making purposes.

Finally, section 4.12.4 described the buffer values that are used as thresholds to define outliers. The higher the target percentage, the larger the number of outliers grows. A lower desired percentage is more easily reached and the buffering procedure is more likely to stop after having examined the distances between objects only, reducing the cases where the buffer is extended to cover the different road representation (length) or data matching errors.

8.5. Text similarity thresholds

Text similarity thresholds were used in data matching stage 3 (65%, justified in section 4.8.3), data matching stage 6 (75%, justified in section 4.8.8) and attribute accuracy stages 2 and 3 (70%, justified in section 4.11.2). The provided justifications, however, do not include some short of sensitivity analysis. This would mean to run the process for an urban area (where there are more chances to find name attributes in VGI sources) multiple times, using different thresholds. For each threshold, a manual evaluation of the attribute accuracy assessment (as sections 4.14 and 5.4.3 described) would then be necessary, and in the end a comparison of the error levels would provide the optimised percentage. A similar analysis is not performed in this thesis, since different sources may still need different thresholds, as proven in the Haiti case study (Tables 7.8 and 7.9).

Attribute accuracy evaluation provides an indication of how efficient the text similarity threshold is (defined at 70% in section 4.11.2). As section 4.14 explained, error type 2 refers to different names mistakenly accepted as similar, meaning that the 70% threshold is quite low. Error type 3, on the other hand, refers to similar names mistakenly rejected as different, meaning that the 70% threshold is quite high. Table 8.9 provides the average values of reference and VGI attribute error percentages, as described in Tables 5.7, 5.8, 6.18 and 7.8. The Haiti cases (case study 3) show that text similarity is not as successful due to far different naming and abbreviations (some examples were presented in Table 7.9). Generally, cases where the threshold seems high prevail, which means that the selected 70% threshold value is quite conservative, and a lower one might have been more efficient.

	Case study 1 (urban area)	**Case study 1 (rural area)**	**Case study 2 (England & Wales)**	**Case study 3 (UN-GMM)**	**Case study 3 (GMM-OSM)**
Error type 2 (low threshold)	0.08%	0.00%	0.04%	0.00%	0.00%
Error type 3 (high threshold)	0.31%	0.81%	0.19%	5.04%	3.91%

Table 8.9: *VGI and reference segment length statistics (m) for the first case study*

8.6. Other issues related to the performance of the method

The method developed is described as automatic, however some manual intervention is necessary. Section 8.6.1 discusses the necessary tasks that need to be applied manually. Section 8.6.2 discusses the performance of the method in computational terms.

8.6.1. Automation and necessary manual intervention

Looking at the flow diagram of the developed method (Figure 4.1), manual tasks are described with dashed lines. These are 'data preparation' and 'evaluation' of the method. The latter was technically described in sections 5.4.2 to 5.4.4.

Data preparation refers to collecting data from the data sources, reprojecting one or both of them to ensure they are in the same Cartesian[8] Reference System and collecting or creating the necessary tessellation file, also in the same Reference System with the datasets. Additionally, the reference dataset needs to have specific column names, which is done either by renaming the columns or by copying data into an empty template that has the necessary table structure. Topology may also need to be corrected, depending on the datasets used. GIS software applications usually provide the necessary tools for an automated topological correction, however this needs to be decided by the user and performed separately. The next step is to load the data (two datasets and tessellation file) into a PostGIS database. There are tools to perform this task automatically, provided that that user has privileges to connect to the database.

Appendix A describes the application developed in PHP to carry out the whole procedure. Briefly, the user connects to the database and selects the necessary files (reference, VGI, tessellation) from a drop-down list. The application then checks that the three data inputs are in the same Reference System and that they overlap. In case of partial overlap, all datasets are clipped accordingly by using bounding boxes, so that only the intersected bounding box (area with data from all inputs) is further processed. The user can then decide to exclude specific road types from reference or / and VGI dataset, however the default and suggested choice (based on the results of this research) is to use all of them. Finally, the user selects the parameters for the positional accuracy evaluation, namely the initial buffer width and the desired level of confidence (in other words the overlap percentage).

[8] A Cartesian Reference System has metric units (e.g. meters), in contrast to a Geographical Reference System where angular units are used (e.g. degrees).

The process begins, following the flow diagram (Figure 4.1). The user waits for the process to finish, being informed through a periodically refreshing web-page on the level of completion (number of processed tiles, compared to their total number). When this is done, the output shapefiles and CSV files (see Appendix A for a complete description) are automatically exported and saved in a folder. Along with them are two template or project files, an ArcMap and a QGIS document, which have the output files pre-loaded, arranged and appropriately symbolised. The user can visualise the results through these project files. It may be necessary to change the project file reference system (because templates use the British National Grid), and to zoom to the extents of the data.

It is then essential to study the results and perform a manual evaluation of data matching, if needed, for a sample of data. Finally, if during the data preparation stage the datasets' structures had to be modified for the procedure, the user may have to revert to the initial structure, if necessary.

8.6.2. Performance in computational terms

The developed method is an automated computational procedure that relies on the capabilities of the hosting machine. Software and hardware characteristics are rapidly improved by technology advances, so the information provided in this section is only to be used as an indication, as the desktop computers used for this thesis have already become obsolete.

For case studies 1 and 3 (Chapters 5 and 7) and part of case study 2 (Chapter 6), the desktop PC used had a 2.66 GHz Dual Core Processor, 4 GB RAM, Windows 7 32-Bit. The whole process for the rural area of Chapter 5, as well as for the Haiti area of Chapter 7, took around 15 hours. The urban area of Greater London needed around 30 hours. The reason for this difference is the increased network complexity in the urban area, which computationally is harder to process.

For most part of case study 2, a faster desktop PC was used (3.20 GHz Core i5 Processor, 8 GB RAM, Windows 7 64-Bit). England and Wales areas were examined region by region, needing up to 3 or 4 days for large regions with dense road network. Reprocessing the Greater London region enabled the comparison of the two systems. The second system needed approximately 18 hours, which shows that as technology advances, the time needed for such an analysis is reduced. On the other hand, VGI datasets become richer with time, which may slow down the procedure in some areas, however Greater London can be safely considered as one of the upper limits of VGI density.

It was noticed that computational time increases in a non-linear way during the processing for both systems, e.g. if 50% of an area finishes in 4 hours, the rest of it will take more than 6 hours, even if the road network has the same density all over the area. A region that was finished in 3 days when using the second system needed more than a week for the first one. The conclusion is that areas need to be divided into smaller ones, depending on their size and the system capabilities. The relevance of the system, however, can also be implied or negated by the application. If the method is to be used for crisis management, computational time is of great importance. However, such cases usually refer to rather limited areas, which could be processed within reasonable time using hardware of average cost. On the other hand, NMAs that deal with national datasets may not have to face a similar time pressure in their production process. In any case, probably their systems are much more advanced and capable of more complicated computations.

8.7. Implications on VGI

8.7.1. VGI and standardization

Among the two VGI sources used so far, there are some differences in the use of standards. So far OSM permitted the schema alteration by users, offering suggestions instead of standards. Only recently, 'Potlatch 2' editor replaced its predecessor and now data types are selected from a range of domain values. Google Map Maker, on the other hand, uses some standards and offers a default schema, where many fields accept information from a range of predefined values (e.g. road types), restricting the user but ensuring database integrity. In this way Google Map Maker offers a specific range of information, while OSM is open to new types, depending on its data contributions.

For the Haiti area where there is access to both the VGI datasets, Chapter 7 showed that OSM is much richer in contribution. Considering that both datasets were strongly enriched quite rapidly after the 2010 disastrous earthquake for humanitarian reasons, the question is why contributors should prefer OSM than Google Map Maker. Google is a well-known web-mapping provider, and if not better known than OSM back in 2010, it could be assumed of a similar reputation. This OSM preference could be either attributed to a more dedicated user group, or to the restrictions posed by Google Map Maker on data collection through the predefined schema, which may have driven new users away and directed them to the next option of OSM. Such a negative effect of the use of standards on VGI was already mentioned as a possibility (Haklay and Weber, 2008). The findings of case study 3 could be in agreement and explain why OSM is richer in Haiti, however more such cases need to be examined to prove the validity of this assumption.

Nevertheless, if data capture methods were standardised in VGI projects, data could be more valuable. If, for example, OSM used a range of values for road types, there would not be 180 different road types for the UK or 33 for Port-au-Prince, among which there are erroneous, misspelled values or even values in different language. Having to choose between a set of values would possibly make the user think twice before appointing a value, instead of using a new expression for an already defined road type. VGI network classification would then be more trustworthy and would enable a more efficient road-type by road-type evaluation, as well as a possible road type rejection from a quality analysis, depending on the reference dataset road types.

Another way of standardisation would be to impose restrictions such as length limits on the secondary road name, which is usually the internationally or nationally agreed conventional road naming (road reference code). This would prevent someone of adding a primary name instead or other irrelevant information, as section 6.5.2 described.

8.7.2. VGI and official data sources

Leaving Wales aside, England shows an impressive level of VGI quality. Data completeness is high in urban areas, reduced in some rural areas, while there are also smaller non-mapped areas in North, Severn and East Anglia region. Wales is by far the less mapped region. Secondary road names are impressively complete for England and Wales. Positional accuracy is averagely found below 12 m. There are, however, areas of positional accuracy below 5 m, as well as areas with much worse results between 40 and 75 m. Additionally, OSM proves to be more updated in few and relatively hard-to-find cases, including objects that should also be present in the ITN dataset. These results are amazing considering the high quality of the ITN dataset and that the OSM project started in the summer of 2004. However, it cannot be suggested that generally OSM can replace the ITN dataset. Nevertheless, some urban areas may provide similar information and could replace reference sources for specific purposes, provided that the positional accuracy is found adequate for the destined usage and scale.

The Haiti case is a different example. Official and VGI sources needed to be rapidly updated in order to be used by the rescue missions after the 2010 earthquake. The UN reference dataset seems to have been collected around 2003, but not updated since then (Figure 7.15). OSM and Google Map Maker, as VGI or crowd-sourced data, were actually collected after the earthquake. This is a good example of how rapidly volunteers can collect and update a spatial dataset in crisis situations, while official sources remain not updated and probably not useful if massive road network changes occur.

Both VGI sources, using the two biggest providers of worldwide satellite imagery (Google and Yahoo!), seem to agree on the position of roads in areas where the UN dataset seems mislocated by more than 100 m (Figure 7.14), which implies a better VGI positional accuracy there. Both VGI sources are more complete in data, with OSM being the richest dataset in features and attributes. As a result, this is a case where VGI could effectively replace official data.

Results also show that although VGI datasets may locally present some level of homogeneity, in general they are heterogeneous datasets. The fact that in many areas there are additional data not gathered by the reference dataset makes VGI sources unique, with their own identity and perspective. Depending on the purpose of usage and quality demands, in such cases they could substitute official sources, which otherwise would have to be purchased and enhanced to cover the special needs of the user (e.g. Figure 6.21).

What also needs to be considered is the efficiency of using a reference dataset for quality evaluation, especially in cases where its quality is lower than the VGI under examination. For example, a generalized official dataset (derived from digitization of 1:100,000 raster maps or even smaller scales), would probably include a limited number of simplified features with lower positional accuracy, due to the scale limitations. For the same area and data types, in contrast, a VGI source permits digitization of equivalent scales of 1:5,000 or even larger by using satellite imagery as a background, which leads to an increased number of more detailed and accurate features. A comparison between such datasets would represent their differences as lower spatial quality of the VGI source, since the reference is considered as the 'ground truth' dataset, which is not what happens in reality. For this reason, it is essential to choose carefully the dataset that will be used as a reference one, based on its quality specifications or metadata. Haklay (2010c), for example, takes into consideration that the OS's 'Meridian' dataset is generalised and within 20 m from the real world position, which leads him to further assumptions about road lengths. This, however, is far more difficult to address when comparing two different VGI sources, since quality is usually unknown and heterogeneous for both. This is why section 7.5.1 describes the comparison results as 'level of agreement' instead of 'completeness' or 'accuracy'.

8.8. Limitations of the method

The method that was developed in this thesis is found to have the following limitations, which provide scope for improvements and further developments:

1. The manual examination in all case studies found cases of objects that although close to each other and of similar size and orientation, they should not be matched. As an example, each dataset (A and B) may have a vertical road (R_A and R_B respectively) to a commonly described road AB (hence matched). R_A and R_B meet road AB at two nearby junctions (first constraint passed) creating a 'T' intersection, they have a similar size and orientation (second and third constraints passed), but they lie on different sides of the road. Although they describe two different roads, they will be erroneously considered as corresponding ones because they both agree to all three geometric constraints, which suggest that in such cases the geometric constrains of stage 1 are not enough. As an example, in Figure 8.9, Features I1 and I2 are correctly matched to O1 and O2 respectively. O3, however, is erroneously considered as a matched feature by being linked to I1. Similarly, I4 is erroneously considered as a matched feature by being linked to O2. This type of error increases data matching percentages (resulting to a more optimistic data completeness estimation) and may increase buffer width (resulting to a more pessimistic positional accuracy estimation, e.g. feature O3 of Figure 8.9). However data matching error levels are low in all study areas because such cases are few and usually refer to short road features.

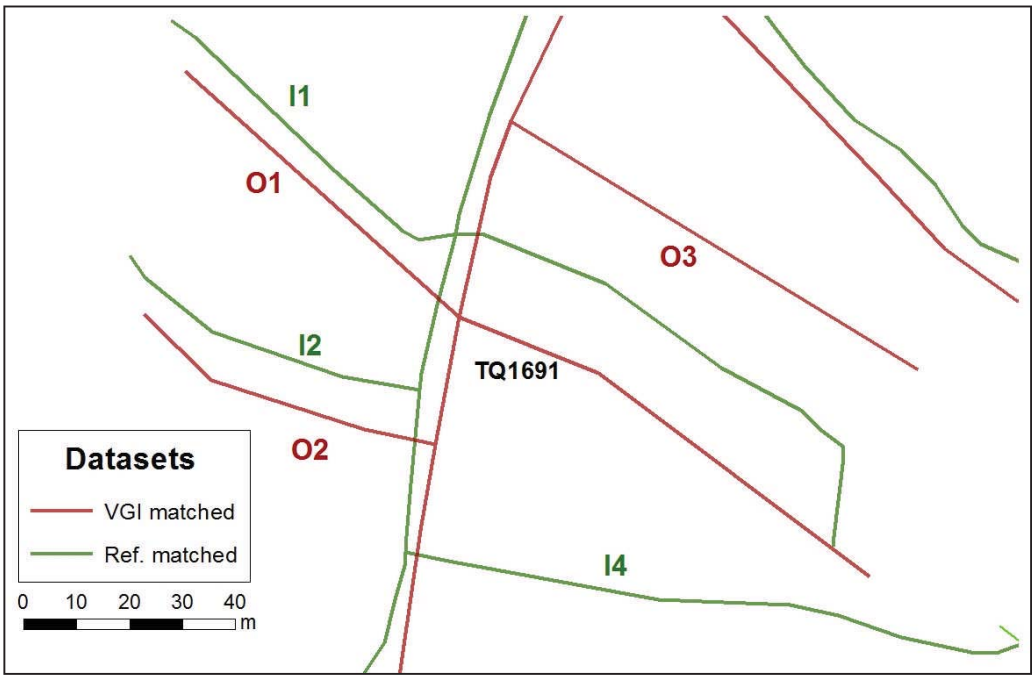

Figure 8.9: *Errors in data matching due to insufficient geometric constraints of stage 1*

2. Since data matching is performed at feature level, features that have different representation (e.g. much longer in one dataset) cannot be further divided to their corresponding part. As a result, data matching is ambiguous in many cases of different representation (Figure 7.13).

Splitting features into smaller parts would provide a better data matching, however there is no profound splitting point: if natural vertices were to be used, this ends up to segments comparison, already performed in stages 1 to 4. However, results showed that this generally is not adequate, especially in Haiti area where successive stages at a feature level contribute significantly to the total data matching. Alternatively, customised splitting points could be based on a buffer and its intersection point, however results would be different depending on the buffer selection, so the method would not provide consistent results.

3. The suggested framework assumes a similar scale size of datasets. If one dataset is derived from a smaller scale map or is somehow generalised, this will lead to bigger and fewer objects in the generalised dataset representing the same data, using a less detailed polyline: generalisation algorithms usually remove vertices from a polyline, creating a smoother shape, which decreases the length of the generalised features, modifies their orientation and lowers their positional accuracy, since they may be partially or totally moved. This increases the distances between corresponding segments, as well as their angular differences. Under such circumstances, the data matching thresholds used in this study (search distances, angular tolerances, length) will not suffice, corresponding objects will fail to be matched and data matching error levels will increase, affecting the measurement of spatial data quality elements accordingly.
4. For attribute accuracy, section 4.11 described how primary and secondary names (if existing) are compared between datasets. What is not examined, however, is the case of secondary VGI names having the value of primary reference ones and vice versa. This may happen in VGI due to the lack of standards, as some users will not read or comply with the project's suggestions about tagging.
5. Attribute accuracy relies on text similarity, comparing and counting similar characters. However, this does not take into consideration the sequence of characters, so roads ABCDE and EDCBA would be considered similar. Although this theoretically seems to be a limitation, in practice such cases are unlikely to happen when examining road names within a certain area. A minor misplacement of letters could occur as a result of misspelling or typing error in VGI, however the road name should still be recognizable.
6. Text similarity (as well as any other matching algorithm for attributes) relies on characters that are understood by the operating system. Before examining datasets in a different language, the language pack has to be installed, otherwise unusual characters will be used for road names. Such case is the 4rth row of Table 4.1 or last row of Table 7.9.
7. Positional accuracy is calculated for a whole tile and not for each feature individually, unlike attribute accuracy or data completeness. As a result, mismatched or differently represented

data may affect positional accuracy value for a whole tile by providing a higher buffer width (lower positional accuracy). Similarly to section 4.8.3, this could be considered as a false negative or type II error that leads to safer choices, since positional accuracy will not be found higher than in reality. Although a threshold is used to define 'outliers' (section 4.12.4) so that such tiles are either not considered for any further use, or serve as a quality measure to correct such errors, tiles that their buffer does not reach the threshold value may also be affected, which leads to a slightly bigger buffer width or more pessimistic positional accuracy results.

8. Although positional accuracy results refer to tiles, values are calculated using matched data only. Non-matched data do not participate in the calculation of the buffer width, so theoretically they may not have the same positional accuracy as other data of the same tile. In practice, however, data collected by a user refer to all types of information within a certain area. The same means are likely to have been used and there is no reason for a quality change according to road types not collected by the reference dataset or close to the tile borders, since both of these factors are not known to VGI contributors. As a result, positional accuracy of VGI matched data within a tile can safely be assumed to apply to non-matched data as well.
9. Section 4.13.2 explained how VGI commission indication relies on road type correspondence, e.g. if reference road type A primarily corresponds to VGI road type B, non-matched VGI roads of type B gain an indication for commissioned data. Choosing the prevailing road type, however, is not as efficient when the corresponding road types could be two or more with close values, for example when the chances of road A corresponding to B and C are 39% and 38%, as opposed to 85% and 13% respectively.
10. Direction of features is not examined, so this method cannot be used to provide information on VGI datasets that need to be used for routing purposes. Additionally, although not tested, the method is unlikely to provide results when the involved datasets have a third dimension (e.g. height).
11. Topology, as defined in this context (section 4.8.3), is not examined or corrected during the automated procedure. Moving from case study 1 to 3, however, its significance grows accordingly. In Chapter 5 there are minor issues of incorrect VGI topology, which do not affect the evaluation. In Chapter 6 there are more such issues, and problems appear on data matching, nevertheless the error levels remain low. In Chapter 7, however, incorrect topology is present in both reference and VGI datasets and its correction is essential for the efficiency of the method. These findings suggest that topology correction should also be automated and integrated instead of being performed when necessary during the data preparation.

12. Section 8.6.1 mentioned that, as part of data preparation, datasets need to follow a specific data structure. This demands a certain level of knowledge regarding GIS software usage. However, the method is designed to be used by someone with such a basic knowledge, as some output files need to be loaded on any GIS software to be visualized, so it is assumed that the mentioned data preparation steps are quite simple.
13. One final limitation is that the application described in Appendix A does not include any visualization interface for the user to monitor the process or view the results. This however is related to the previous issue, as some basic knowledge of GIS is enough to access the database through GIS software (e.g. QGIS) for monitoring the process (see Appendix A - Figure 9) or afterwards for viewing, inspecting and analyzing the results.

8.9. Potential usage of this research

This section provides some examples of who could benefit from this research, describing also some existing cases of VGI usage and linking back to what the introductory Chapter 1 discussed.

8.9.1. National Mapping Agencies (NMAs) already using crowd-sourced information

United States Geological Survey (USGS) is maybe the oldest example of VGI usage from an official cartographic organisation. Their 'National Map Corps' program started back in 1992 by asking volunteers to identify topographic map errors and suggest corrections. However, due to the volume of changes, it was not possible to evaluate and apply them in a reasonable timeframe. This drove users away, resulting in the program's failure not because of the users' immaturity, but because of the lack of the appropriate tools made available to the professionals. In 2003 they brought it back to life, and since 2007 it supports GPS and web-based procedures (Bearden, 2007). However, due to the volume of data that needs evaluation, GPS procedures stopped in 1 August 2008 (National Map Corps, 2010), probably to avoid a similar failure. Now it is a pilot program, where users need to register. They also need to follow standards regarding which features will be mapped and how they will be tagged. This project bares some similarities to OSM, although the variety of data stored is much more restricted. One of their goals is to automate the verification and processing of crowd-sourced data (National Map Corps, 2011), which will allow them to re-enable the GPS procedures. An automated quality evaluation similar to the proposed in this research could help them reach their goals. Another **USGS** example described by Bearden (2007) is also a National Maps Corps project where volunteers are invited to collect map-worthy structures using GPS receivers and a proper tagging.

The Federal Office of Topography **Swisstopo** (Switzerland) updates and improves the Topographic Landscape Model of the country using revision notifications by volunteers through a web browser and a one-way workflow. Crowdsourced data are automatically integrated in a revision layer every night, and then they are assigned to the corresponding line of production for each product to be further processed and evaluated. Upload of GPS tracks is also sought to be enabled in the future (Guélat, 2009).

UK's **Ordnance Survey (OS)** invests in research in Vernacular Geography (Ordnance Survey, 2011). Vernacular geography is the perception of places in ordinary people's language, which is something that an NMA should consider when labelling its spatial products. OS uses volunteers to collect place names (People's Place Names, 2010). The user first defines the area (giving a postal code or pointing on a map) and then gives related place names (area name, nearby landmarks, etc), while at the same time the user provides information about age and familiarity with the area, to be used as an indirect credibility measurement. Users can also add photos, view their place names on the map (or those of others) or discuss place names. This has the form of a survey, with no user registration necessary, but with the bait of a potential prize if user identification is given. A future analysis of an adequate volume of data could improve existing gazetteers, putting into effect Goodchild's (2008b) idea of using indigenous experience in local naming information to enrich and update gazetteers, which so far NMAs did not succeed in collecting or representing. The proposed framework can help in this direction and can be considered as a different collection method for Vernacular Geography, since the VGI features found as non-accurate in terms of attribute accuracy may indicate a local name value that differs from the official one.

Another project of **OS** that uses crowd-sourcing is 'explore' (Ordnance Survey, 2010b), in which users can create and share routes in an area, including route description, POIs or photos. User account creation is necessary. After plotting the route, the user is asked to identify the means (walking, biking, flying, sailing, etc), so a crowd-sourced database is created, containing features that traditionally have not been mapped. The provided framework can help in this direction, as it directly extracts such information from VGI sources.

Finally, **OS**'s project 'GeoUsers', is in seek of understanding user needs and how they will form in the future (Ordnance Survey, 2010a). This is done by collecting non-spatial information to improve their products range: by asking users to describe their spatial needs, they can redirect their production policies, in order to produce spatial data that will fulfil their customers' spatial requirements in a

better way and at the same time increase their income. By examining the volume of specific non-matched VGI road types provided by the proposed framework, a new user demand may be implied, which renders this research valuable in this area as well.

8.9.2. Commercial Mapping Organisations (MOs) already using crowd-sourced information

Google is a large commercial organisation that shifted from official data to VGI (Helft, 2009). By dropping commercial data, it relies on volunteers to update spatial data in about 140 countries, saving money and at the same time providing a more frequently updated service. This seems to put into effect Goodchild's notion of 'using citizens as voluntary sensors' (Goodchild, 2007a).

TeleAtlas updates its road network datasets using feedback from TomTom's navigation devices that citizens possess (Helft, 2009). The new data are integrated and provided back to the users for their navigation purposes with an indication regarding their confirmation status, so that each user can choose whether to use them or not (Mac Gillavry, 2009).

In both examples, data collected by volunteers need to be evaluated before being integrated in the final spatial product, so the proposed framework is also applicable.

8.9.3. Opportunities for NMAs and other commercial MOs

Results showed that the method can efficiently isolate data missing from the reference dataset, which is the first step of conflation. All case studies provided a non-matched VGI dataset as output, which includes all the data not present in the official dataset. In this way reference datasets could be updated or even enriched with additional types of information not collected so far, putting into effect what is discussed in sections 1.5 and 1.8. Positional accuracy results can ensure that the examined source will not reduce the quality of the new product.

The above, however, assume that there is no legal framework in VGI source that would prohibit such integration. Considering the purpose of usage (e.g. production of new datasets for commercial reasons) and the VGI source terms and conditions, a closer look needs to ensure that no copyright issues are violated. OSM's license agreement, for example, allows for its free data to be used for commercial reasons, but at the same they have to be free for re-usage, preventing someone to claim their ownership. This may not come into terms with the NMA's or MO's copyright framework regarding the new product.

Section 1.8 suggested that VGI could be used to limit NMA updating procedures, time and cost by helping prioritize the areas to be processed. This is a faster way for an organization to keep its important data up-to-date. The proposed methodology could help in that direction through an examination of the VGI non-matched output dataset, which includes data not present in the official dataset. The priority analysis could consider the road types and their importance, the amount of missing data in an area, its population and other aspects that make this place important, such as touristic attractions or commercial activities.

Positional accuracy values, combined with data completeness, could also help NMAs or MOs to protect their copyrights. Tiles with reference and VGI matched percentages close to or 100% at the same time, also with buffer widths close to zero, indicate extremely high data agreement and positional accuracy. Considering the complex way that data completeness and positional accuracy are calculated, this looks rather suspicious and may imply that a VGI user has contributed copyrighted data to the VGI source. In this case, legal actions may need to be taken against the VGI source. Although a professional data source may already use 'Copyright Easter Eggs'[9] to check for copyright theft, this is a different tool which applies to the whole area, with no need to intentionally falsify some data.

Section 1.8 mentioned that VGI could be used for propaganda or boundary disputes, so an NMA should be vigilant. In this context, this research can be used to check for deliberate data distortions. Cases of different language used for attributes can be spotted by non-accurate matched VGI data inside and around the areas of dispute. Data completeness can show if data extend to the national border line, or part of them is intentionally attributed to the neighboring country.

NMAs and MOs invest many resources on updating their datasets, hoping that they will be compensated when selling their spatial products. Therefore, it is essential to know how inferior or superior their products are, so that they can act accordingly to improve their datasets or demonstrate their superiority and establish their sales and reputation.

8.9.4. Disaster management

Section 1.5 mentioned cases of VGI usage after natural disasters. The third case study in Haiti is one example, which shows a high level of dedication and determination of VGI contributors, who proved

[9] A 'Copyright Easter Egg', in terms of mapping, is an intentional error in data. It can be a non-existing map feature, a distorted one, an erroneous or misspelled name etc. Its purpose is to help identify its original author and to prove copyright theft (OpenStreetMap, 2012a).

to be able to rapidly update or create datasets much faster than an official source (Figure 7.15). Although an official source with limited resources and personnel cannot be compared to the number of volunteers regarding data updating, they can help by comparing the VGI sources with their data, provide the results and suggest the more suitable data source for the search and rescue teams. The suggested automated method can be applied frequently, since data are collected as long as volunteers exist, providing a dynamic quality image of the available data. This ensures that the best available data are used by the search and rescue teams, which may save time, resources and, hopefully, lives. By this, official sources keep offering valuable services, even when they cannot update their datasets as fast as the crisis situation demands, while they also aid to the coordination of VGI, pointing out where data are missing and volunteers are needed. In this way, collaboration between official sources and VGI leads to better choices for the common good.

There will be, eventually, cases with no or poor official data. Chapter 7 showed that this method can also be used to compare two VGI sources, considering one of them as reference. In case that no NMA exists to run this comparison procedure, the engineering department of an involved administrative authority should be able to do it.

8.9.5. Governmental, non-governmental organizations and VGI projects

Governmental organizations, as well as some non-governmental ones, usually have access to official spatial datasets at no or low cost. However, their needs for spatial data may not be fully covered by the standardized types of information offered by an NMA or MO. Environmental organizations need also data types closer to nature, such as footpaths or trails on a mountain or inside a forest. Such data cover a market group of insignificant size for an NMA or MO and they are difficult to collect. The Adaptable Suburbs research project mentioned in section 1.5 (EPSRC, 2011), for example, needs to combine the additional information on footpaths, not included in the official dataset, with the existing official data for pedestrians. These data can be easily collected by the VGI non-matched output dataset of the proposed framework.

There are cases of non-governmental organizations with no free access to official spatial sources and the cost may be sometimes forbidding for their budget. In the context of this research, there are two options. One is to gain access to the results of such an analysis, if performed by another organization. The second is to use this research to compare two VGI sources, as described in Chapter 7, considering one of them as reference source. In this way they can either select the more suitable VGI source, or combine and conflate them.

Developing countries can also benefit in two ways. They can compare and combine VGI datasets in their area if no official datasets are available. If there are official data, they can compare them with VGI datasets and examine how reliable and up-to-date the official or VGI datasets are, which will enable a more efficient decision-planning and infrastructure services provision.

VGI projects rely on their dedicated users to expand their coverage and reputation. Their designers could apply this method to compare their dataset with another VGI source, if available. This can give them some indication on the density of their data, and they can direct their contributors to areas of relatively scarce coverage. They will also be able to discover some of the errors in their database, such as the minor ones described in Figure 5.25b (data inconsistency), the major ones in Figure 7.14 (errors in positioning), or the ones in section 6.5.2 (unusual spatial patterns). Contribution among users of a single project is often based on competition and the desire to rise to the top of the contributors' hall of fame. The comparison between different projects will expand the rivalry outside the same project, urging for more contributions and leading to VGI sources of higher quality.

By combining the quality results with the user or users within an area, assumptions can be made for their credibility, especially when the reference dataset is an official one. If VGI projects integrate this as a user attribute, competition is likely to become more professional and data collection more careful, improving VGI quality further. This is further discussed as part of future research (section 9.4.1).

8.9.6. Defense mapping

For reasons of defense and strategic planning, it may be essential for one country to have maps of its neighbors. Some basic datasets are usually provided within members of alliances such as NATO, or can be found by other sources. VGI could be a source to provide a more detailed dataset for non-accessible areas. When compared with the existing datasets, this research can help create a more detailed database: new features could be added using the VGI non-matched dataset, while positional accuracy of matched data could be assumed to be the same for the new data within the tile. Since for national security reasons the output products will be classified and not publicly distributed, copyright issues regarding the VGI project license framework will be easier to handle.

8.10. Summary

This chapter argued on some global parameters that were used in all study cases, justifying their selection. It further discussed other issues, regarding the performance and automation level of the proposed method. Based on the results of the three previous chapters, implications of the method on VGI were discussed.

The method was similarly applied in three case studies and, despite the different results and their interpretation, errors were consistently kept at low levels. This proves the efficiency and robustness of the method. Nevertheless, there are some limitations that needed to be mentioned and briefly described. This chapter finished with suggestions on the usage of this method, partially following up or linking back to the discussion of sections 1.5 and 1.8, and enriching it with a more detailed usage description at the level of output files and results, as well as with additional examples. VGI is already being taken seriously by some professionals, and this thesis aims to facilitate the expansion of VGI usage.

Chapter 9

Conclusion

9. Conclusion

9.1. Introduction

This final chapter concludes the thesis by linking it back to where it started. The Research Aim and Objectives that were set in the beginning are revised and final conclusions and opportunities are discussed. The limitations of this framework, as well as related issues to this research that were not explored, form some suggestions for further research, which close this thesis.

9.2. Meeting the Research Aim and Objectives

Chapter 1 stated the Research Aim of developing a framework for the quality evaluation of VGI. This needed to be materialized by an automated method that would include the necessary steps to evaluate the basic aspects of VGI quality and would be appropriately designed to deal with its nature, which is far different from that of existing standardized official datasets. The automation is essential to enable the method repetition in the future, as well as its usage in relatively large areas. This section discusses how the Research Objectives, set in section 4.3 to reach the Research Aim, were met in the context of this research.

9.2.1. Understand the nature of VGI linear data

Defining the general characteristics of a VGI linear dataset that could be used in a comparison process regardless of the source is quite difficult, since each data source would probably have a different structure. By choosing the feature as the basic object unit, despite the second limitation mentioned in section 8.8, the method worked similarly in all tested areas, even when different data sources were used.

The significance of spatial and non-spatial attributes is flexible. All objects have geometric attributes, but not all of them thematic ones. Wherever present, however, they need to be taken into consideration and be used to improve the quality analysis. Primary and secondary names were chosen as the non-spatial attributes that are likely to exist in all VGI sources, however cases of sources with not even these had to be predicted, so that the quality analysis could be applicable in general, even in such cases. A third attribute referring to the road type information, although more likely to exist in all VGI sources, is not compared due to the different classification followed by each

dataset. Nevertheless, it is also used during the analysis to provide other types of information (e.g. indicate data commission).

By using tiles to split the data and examine them piece by piece, individual results that are more representative of the local data quality are efficiently provided. The use of the extended tiles (section 4.6) ensures that the analysis for objects close to or crossing the tile borders produces similar results even if the tile position, size or shape is modified. This proves the efficiency of the method to deal with VGI heterogeneity while being robust at the same time.

The spatial data quality elements chosen to be examined for VGI (section 3.2), combined with what was considered significant to be produced by a quality analysis, guided the design of the theoretical model. The need to put it into practice led to the next objectives.

9.2.2. Develop a suitable automated data matching procedure

To achieve the flexibility of examining spatial and non-spatial attributes, an iterative data matching process was developed (section 4.8). By processing data in a seven-part sequence, the significance is shifted to geometric attributes when non-spatial ones are not present in VGI dataset. Despite the differences in attributes' completeness within the same VGI dataset, or even lack of attributes, the evaluation in all case studies showed that data matching maintained similar and low error levels, so the balance between examining spatial and non-spatial attributes is automatically adapted to the input data and proves to be quite effective.

Regarding the different approaches used in data matching, having to process the same data seven times is quite hard in computational terms, however the stages are designed to successively reduce the volume of data that needs to be examined by the next ones. Considering the results of all case studies and the stages' contribution to data matching, as well as data matching errors, the conclusion is that all stages are necessary for an efficient automated object matching with low errors, however their significance will be different depending on the nature of the input reference and VGI datasets.

9.2.3. Perform quality analysis

The automated data matching procedure efficiently isolated data that are present in both datasets in all case studies, preparing the data for the quality analysis. It proved to be a much more significant part of the research than expected, because quality results have to be based on corresponding

objects, so errors in their correspondence are inherited to quality results. Considering the spatial quality elements of the first objective, the data matching errors of the second one and the gaps in the literature regarding the appropriate quality indicators (section 4.2), the indicators chosen in this context provide a more detailed and systematic quality evaluation, compared with what has been offered so far.

By using the length of matched data, compared to the total length for both datasets, the calculated data completeness is independent of the number of features used by each dataset, so it is not affected by the different number and positioning of features used by each input source to represent the same object. Additionally, the use of road types during the examination (described in sections 4.10, 4.13.1 and 4.13.2) allows for further improvement of the analysis, specifically for data completeness evaluation: primarily, road type correspondence is collected, based on feature correspondence, giving an indication of significant non-matched features that could be considered as commissioned data. Apart from what is missing, excess data are also important to be found for a thorough data completeness evaluation. In the case of datasets with different objectives, as in the VGI case, the types of information will be different between the datasets, so road type correspondence is necessary to decide on data completeness and indicate data commission.

Attribute accuracy also relies on feature length and feature correspondence, while the way it is performed partially handles errors of the automated object correspondence. Hence, the resulting value is independent of the number of features which may vary between datasets, it is also independent of the number of distinct names which would provide erroneous values in cases of misspelling, and, finally, it includes attribute completeness as well by considering partially missing attributes (e.g. when only one of two VGI features representing the same road has a name) (section 4.11.1).

For positional accuracy, a method already tested in its simplified form by others is selected. However, this research innovatively uses it in its advanced version, which produces an accuracy value for a desired level of confidence (section 4.12). Although this has been the recommended approach of applying the positional accuracy method since its definition back in 1997 (as section 4.12.1 described), this seems to be the first study that applies it.

9.3. Final conclusions and opportunities

By meeting the Research Objectives, the Research Aim is successfully accomplished. Quality results succeed in representing VGI heterogeneity by referring to tiles, however the detailed level of examination to produce them provides more accurate quality estimations and is only possible due to the efficient way of finding corresponding objects between datasets. Further down to the feature level, each feature is marked if found matched or accurate in terms of attributes. Information is also stored for the attribute accuracy type (exact or similar name matching).

This is, arguably, the first study in VGI that includes an efficient data matching and successive estimation of the three spatial data quality elements (data completeness, attribute and positional accuracy), which this thesis considers as most important for VGI.

The range of applications of the provided framework extends those implied by the research objectives. Apart from the output values that describe VGI quality locally, this framework offers additional opportunities. Although sporadically discussed in previous chapters when met, they are collectively presented thereafter.

Output dataset of non-matched features can be used for data fusion purposes in order to enhance one of the two datasets with data from the other. This enables this framework to be partially used in the different research area of 'conflation' (described in section 3.3.1). The detailed approach for object correspondence in this research can serve as the first of the two general stages of conflation.

Road type examination can additionally provide new opportunities. One output table provides information of what is generally mapped and what is not by each dataset, based on matched and non-matched road types (sections 5.5.3 and 7.5.3). This helps improving data collection to cover non-mapped road types, e.g. OSM users could pay more attention not to miss ITN's 'Alleys' in the UK. Another use of this road type correspondence is to find out which types of information are generally unique in one dataset, so they could be removed to create two more homogeneous datasets in terms of the information they provide. This would justify a possible road type selection to perform the analysis, although section 5.5.3 argued against it. Automation makes it possible to re-run the evaluation and compare the results.

This framework can additionally be applied on two different VGI sources to compare and evaluate one against the other in the same way as with official datasets. This expands its usage to those who

cannot afford the official data, do not have access or when official data simply do not exist, as well as to VGI project designers who could compare their project with others.

This thesis follows the direction shown by researchers that were mentioned in Chapter 2 (Goodchild, 2007a; Sieber, 2007; Boin and Hunter, 2007; Goodchild, 2008a; Goodchild, 2008b; Haklay and Weber, 2008; Flanagin and Metzger, 2008; Coote and Rackham, 2008; Maué and Schade, 2008; Auer and Zipf, 2009; Elwood, 2009; Haklay, 2010c, Antoniou *et al.*, 2010a). They all realize that VGI is heterogeneous; most of them suggest that new methods need to be implemented to deal with VGI heterogeneity and assess its unknown quality, while some of them move further to provide theoretical or practical approaches. This thesis proposes a practical approach and, with the use of suitable metrics, specific quality elements are calculated to describe spatial data quality of linear datasets.

As Chapter 2 showed, VGI is an expanding trend with many aspects: various applications collect different types of spatial information for diverse purposes. Hence, it is quite difficult for a quality approach to be suitable to tackle all or many of these aspects at the same time. In this context, this research tackles VGI heterogeneity and quality assessment in an automatic and systematic way, however it is confined to the linear datasets.

9.4. Suggested further research

This research seems to have succeeded in providing a methodology to answer specific questions on VGI quality, however new ones were raised and point out directions for further research. Additionally, limitations or implications imply future work to improve the described framework.

9.4.1. New directions

One direction is to combine the quality results with the users' information and assess their credibility. This is useful to evaluate data where there are no official datasets available. Additionally, notifying the users of their data quality would also improve their performance, boost their competitiveness and lead to VGI of higher quality. In this context, the described method could be slightly altered to provide such information. User ID is not provided in the shapefile format that was used, however it can be extracted from the OSM xml file for each feature and become a new attribute in the shapefile. Combined with the results of data matching, it would provide information on the amount of new data that each user offers (non-matched VGI), while it would also serve as a

confidence level for further results (matched VGI), e.g. if 83% of the data contributed by user 'A' was found matching, results regarding attribute and positional accuracy would only refer to 83% of user 'A' data. Each user's matched data could then be examined separately, regardless of the tiles already used for data matching. Using a buffer wider than the maximum search distance, a new tile could be created for each user (of variable size and shape). Matched reference data within this new tile could then be compared and evaluated for attribute and positional accuracy, providing quality results per user. Within the same wider area, it can safely be assumed that the results would generally apply to the user's non-matched data as well, providing another indirect method to evaluate the whole VGI source (and not only what corresponds to the reference dataset).

Since the analysis of a wider area is now possible, as shown in Chapter 6, it would be interesting to apply geostatistical tools to study further quality correlation, for example if positional accuracy is related to attribute accuracy or data completeness. Some spatial patterns were found in this research in favor and some against an assumption of correlated quality elements. Spatial regression is an appropriate geostatistical method that can provide a deeper insight and test such hypotheses.

Chapter 7 experimented by applying the method on two different VGI sources, and results showed that it performs similarly efficiently. This adds a new dimension to the scope of the analysis, providing a way to evaluate a VGI source when no reference dataset is available. This requires a decision on which dataset should be considered as reference, because the positional accuracy approach will use it to apply the buffer. However, although the results are produced in the same way, they could be named differently, for example positional accuracy would refer to the distance between the datasets, as none could be considered as accurate; accordingly, 'agreement' would be a more appropriate term instead of 'completeness', when following the terms 'data' and 'attribute'. In this context, the use of Buffer Overlay Statistics (Tveite and Langaas, 1999) might be more appropriate for the positional accuracy evaluation than the selected Increasing Buffer Method (Goodchild and Hunter, 1997), because it does not assume differences in quality (section 3.3.4). This needs to be tested and the method needs to be applied on further cases where more than one VGI sources are available.

Additionally, this framework could be tested if it could also be applied on two reference datasets. Each reference dataset is assumed to be homogeneous, however there may be differences between them in density, positional and attribute accuracy. The provided approach could find these differences and test this assumption, e.g. varying values of data completeness, positional or attribute accuracy would imply that one or both datasets are not as homogeneous as they claim to

be. However, some modifications regarding attribute accuracy are essential: while for VGI misspelling was excused, official datasets are not expected to have such errors. The use of abbreviations may be a problem, however it is usually standardized (unlike in VGI). This may enable a different approach than addressing text similarity based on characters, which would also deal with the fifth limitation of section 8.8.

While this thesis is focused on linear datasets, it is applied on road networks only, due to their importance. It needs to be tested if it is applicable to other linear datasets, such as water or power networks. Additionally, VGI offers other data types that demand a different methodology regarding their quality. Such data types are points (e.g. various points of interest) and polygons (e.g. land use, buildings). This research could be combined with other existing or future quality approaches for these data types and form a suite for a complete evaluation of a VGI data source, which would include all the provided data types.

While this thesis focused on defining the elements of VGI quality and the establishment of metrics to measure them, there is a need to communicate the results appropriately. This is the third part of spatial data quality theory (Servigne *et al.*, 2006), which here is addressed in a rather unsophisticated way, with a series of output files that can be accessed through a suitable software. Visualisation could grant a better insight of the results, providing ways to highlight the important ones and a more professional interface for the application described in Appendix A (thirteenth limitation of section 8.8). Devillers *et al.* (2007) present a 'Multidimensional User Manual' (MUM) prototype to communicate spatial data quality information of heterogeneous datasets. Visualisation refers to spatial data quality elements (one dimension), aggregating and providing results from a feature level to a larger area (other dimension) and offering alternate paths for a detailed analysis. This prototype seems ideal for the detailed level of data quality results that this thesis provides (from feature to tile level).

9.4.2. Future improvements of the framework

There are some limitations in automation which could be rectified. Although differently proven in case study 1, when reaching case study 3 it was made clear that topology needs to be corrected for both datasets when it does not meet the requirements mentioned in section 4.8.3. The automated process needs to include topology correction as well, which would tackle the eleventh limitation of section 8.8. Additionally, the code needs to be enhanced so that no specific data structure has to be followed: the user should be asked to define the data columns that correspond to the necessary

information to perform the analysis, so that the appropriate tables will be automatically created internally after defining the datasets (twelfth limitation).

Regarding data matching and the first limitation of section 8.8, a mutual overlap examination similar to the one described by Gabay and Doytsher (2000) could further reduce data matching errors, by avoiding matching segments when one is located beyond the end of the other during stage 1 of the data matching process (described in section 4.8.3).

Regarding the attribute accuracy approach and the fourth limitation of section 8.8, cases of VGI secondary names having values of primary ones and vice versa need to be taken into consideration. However, this will be quite complicated and will increase the processing time: due to the nature of secondary names (coded national or regional roads, e.g. M25, A204), only exact name matching is examined between the datasets (see section 4.11.3). In order to compare primary with secondary name values, text similarity should also be applied when comparing secondary reference with primary VGI names, as well as primary reference with secondary VGI names, which means that the attribute accuracy algorithm would include four 'full' comparisons (exact matching and text similarity) instead of one 'full' and one 'simple' (exact name matching only). Useful information can be found in Ludwig *et al.* (2010), who do this cross-examination using the Levenshtein algorithm (see section 3.4).

Regarding the positional accuracy approach and the seventh limitation, buffering could be applied individually for each VGI feature, so that positional accuracy is calculated at feature level. The tile result could then be derived from a statistical examination of all features within the tile, which might deal with outliers more efficiently and produce a more optimistic tile value. Additionally, each feature would have its own positional accuracy, which brings the evaluation to the feature level similarly to the other quality elements examined. Processing time is likely to increase, however due to the increasing hardware capabilities this will soon not be a concern, although a further investigation of the algorithm's performance is needed.

Some of the parameter values that were presented in Tables 4.7 and 4.8, were decided using trial-and-error methods or other tests, based on sample data of the case studies. Although manual evaluations imply that they are efficiently chosen, it needs to be examined if and for which of them a sensitivity analysis can be performed. This will ensure that they are optimised, producing in turn

more accurate results. For those, however, that depend on the nature of data sources, such as the text similarity thresholds, the user should be able to interact and modify them accordingly.

Preliminary use of this framework in less accurate datasets, where one of them is generalized, proved that data matching is not as efficient (third limitation). The problem with bigger distances and angular tolerances between corresponding objects when one dataset is generalized can be solved by increasing the GPS-assumed-accuracy parameter 'a', which is used in the calculation of both constraints (section 4.8.3). As a result, the application described in Appendix A needs to include it as a user-defined parameter, so that the user can try different values in a sample area and choose the one that leads to better data matching. Additionally, by enabling the user to modify the parameters of Tables 4.7 and 4.8, it would be easier for the method to be customized according to the data sources that are used as input.

References

References

Agumya, A. and Hunter, G.J., 1999. Translating uncertainty in geographical data into risk in decisions. In W. Shi, M. F. Goodchild, & P. F. Fisher (Eds.), *Proceedings of the International Symposium on Spatial Data Quality*, p. 574–584. Hong Kong, China. 18–20 Jul 1999.

Agumya, A. and Hunter, G.J., 2002. Responding to the consequences of uncertainty in geographical data. *International Journal of Geographical Information Science,* 16(5):405-417.

Al-Bakri, M. and Fairbairn, D., 2010. Assessing the accuracy of 'crowdsourced' data and its integration with official spatial data sets. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, p.318-320. Leicester, UK. 20-23 Jul 2010.

Amazon, 2011. http://www.amazon.com/ [accessed 25 Feb 2010]

Antoniou, V.; Haklay, M., and Morley, J., 2010a. A step towards the improvement of spatial data quality of Web 2.0 geo-applications: the case of OpenStreetMap. In *Proceedings from GIS Research UK 18th Annual Conference.* p. 197-201. University College London, UK. 14-16 Apr 2010.

Antoniou, V.; Morley, J. and Haklay, M., 2010b. Web 2.0 Geotagged Photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1):333-340.

Ariza-López , F.J.; Mozas-Calvache, A.T.; Ureña-Cámara, M.A.; Alba-Fernández, V.; García-Balboa, J.L.; Rodríguez-Avi, J. and Ruiz-Lendínez, J.J., 2011. Influence of sample size on line-based positional assessment methods for road data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(5):708-719.

Ather, A., 2009. *A Quality Analysis Of Openstreetmap Data*. MSc, University College London, UK.

Auer, M. and Zipf, A., 2009. *How do Free and Open Geodata and Open Standards fit together?* In OSGIS (Open Source Geospatial) *1st OSGIS Conference 2009*, Nottingham, UK, 22 Jun 2009.

Basiouka, S., 2009. *Evaluation of the Openstreetmap Quality.* MSc, University College London, UK.

BBC news, 2010. http://news.bbc.co.uk/1/hi/8597779.stm [accessed 22 Apr 2010]

Bearden, M.J., 2007. The National Map Corps. The USGS's Volunteer Geographic Information Program. [Position Paper] In *Workshop on Volunteered Geographic Information*. University of California, Santa Barbara, USA. 13-14 Dec 2007.

Boin, A.T. and Hunter, G.J., 2007. What communicates Quality to the spatial data consumer? In ISSDQ (International Symposium on Spatial Data Quality) *5th International Symposium on Spatial Data Quality 2007,* Enschede, The Netherlands. 12-15 Jun 2007.

Booking, 2010. http://www.booking.com [accessed 25 Feb 2010]

Brotzman, D., 2009. Crowd Sourced Data - Be Very Afraid or Time to Rejoice? *VGIS News,* [Online] November. Available at: http://www.vcgi.org/commres/?page=./publications/default_content.cfm [Accessed 19 Apr 2010]

Brown, J., 2001. Three case studies. *In* Werry, C. and Mowbray, M., *eds.* Online communities: commerce, community action, and the virtual university. Upper Saddle River, NJ: Prentice Hall. Ch 2, *p.*33–46.

Bruns, A., 2008. The future is User-Led: The path towards widespread produsage. *Fibreculture Journal* [Online]. Issue 11. Available at: http://journal.fibreculture.org/issue11/issue11_bruns.html [Accessed 4 Mar 2010]

Budhathoki, N.R.; Bruce, B. and Nedovic-Budic, Z., 2008. Reconceptualizing the role of the user of Spatial Data Infrastructure. *GeoJournal*, 72:149-160.

Budhathoki, N.R.; Nedovic-Budic, Z. and Bruce, B., 2009. A Framework for Understanding Participants' Motivation in Voluntary Contribution of Geographic Information. [slide presentation] In GSDI (Geo Spatial Data Infrastructure), 2009. *GSDI 11 World Conference - Building SDI Bridges to address Global Challenges*, Rotterdam, The Netherlands, 15-19 Jun 2009.

Caprioli, M.; Scognamiglio, A.; Strisciuglio, G. and Tarantino, E., 2003. In *Proceedings of the 21st International Cartographic Conference (ICC) 'Cartographic Renaissance'*. Durban, South Africa. 10-16 Aug 2003.

Charras, C. And Lecroq, T., 1998. Sequence Comparison. Available at http://www-igm.univ-mlv.fr/~lecroq/seqcomp/ [accessed 30 Oct 2011]

Chilton, S., 2009a. Crowdsourcing is radically changing the geodata landscape: Case study of OpenStreetMap. In 24th ICC (International Cartographic Conference). *The World's Geo-Spatial Solutions*. Santiago, Chile. 15-21 Nov 2009.

Chilton, S., 2009b. Data In, Data Out – Openstreetmap incorporates more geodata. In AGI (Association for Geographic Information), 2009. *AGI GeoCommunity Conference 2009 - Realising the value of place.* Stratford, U.K. 22-24 Sep 2009.

Chrisman, N., 1989. Error in Categorical Maps: Testing Versus Simulation. In *Auto-Carto 9: Proceedings of the 9th International Symposium on Computer-Assisted Cartography, ASPRS/ACSM*, p. 521–529. Baltimore, USA.

Chrisman, N., 2006. Development in the Treatment of Spatial Data Quality. In R. Devillers and R. Jeansoulin, eds. *Fundamentals of Spatial Data Quality*. London: ISTE Ltd. Ch. 1, p.21-30.

Cipeluch, B.; Jacob, R.; Winstanley, A. and Mooney, P., 2010. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, p. 337-340. Leicester, UK.

Coleman, D.J. and Georgiadou, Y. and Labonte, J., 2009. Volunteered Geographic Information: The nature and motivation of produsers, *International Journal of Spatial Data Infrastructures Research,* Special Issue GSDI-11.

Coote, A. and Rackham, L., 2008. Neogeographic Data Quality: Is It An Issue? In AGI (Association for Geographic Information), 2008. *AGI GeoCommunity Conference 2008*. Stratford-Upon-Avon, U.K. 24-25 Sep 2008.

Craglia, M., 2007. *Volunteered Geographic Information and Spatial Data Infrastructures: when do parallel lines converge?* [Position Paper for the VGI Specialist Meeting]. Santa Barbara, USA 13-14 Dec 2007.

CRSSA, 2012. PPGIS Examples: Looking for patterns http://www.crssa.rutgers.edu/ppgis/CaseStudies.htm [accessed 30 Jan 2012]

de Bruin, S., 2008. Modelling Positional Uncertainty of Line Features by Accounting for Stochastic Deviations from Straight Line Segments. *Transactions in GIS,* 12(2):165-177.

de Smith, M.J., Goodchild, M.F. and Longley, P.A., 2009. *Geospatial Analysis - a comprehensive guide: Directional analysis of linear datasets*. 3rd ed. Available at: http://www.spatialanalysisonline.com/output/html/Directionalanalysisoflineardatasets.html

Devillers, R. and Jeansoulin, R., 2006. Spatial Data Quality: Concepts. In R. Devillers and R. Jeansoulin, eds. *Fundamentals of Spatial Data Quality*. London: ISTE Ltd. Ch. 2, p.31-42.

Devillers, R.; Bédard, Y. and Jeansoulin, R., 2005. Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use within GIS. *Photogrammetric Engineering & Remote Sensing*, 71(2):205-215.

Devillers, R.; Bédard, Y.; Fisher, P.; Stein, A.; Chrisman, N. and Shi, W., 2010. Thirty Years of Research on Spatial Data Quality: Achievements, Failures and Opportunities. *Transactions in GIS*, 14(4):387-400.

Devillers, R.; Bédard, Y.; Jeansoulin, R. Moulin, B., 2007. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *Transactions in GIS*, 21(3):261-282.

Devogele, T.; Trevisan, J. and Raynal, L., 1996. Building a Multi-Scale Database with Scale-Transition Relationships. in: M. J. Kraak and M. Molenaar (Eds.), *7th International Symposium on Spatial Data Handling*, Delft, The Netherlands, Taylor & Francis, p. 619-633.

DGI, 2010. http://www.wbresearch.com/dgieurope/dayone.aspx [accessed 2 Mar 2010]

Dobson, J.E. and Fisher, P.F., 2003. Geoslavery. *IEEE Technology and Society Magazine*, 22(1):47-53.

Dobson, J.E., 2008. *Geoslavery in the Stacy Peterson case.* [Online]. Document Resources for Small Business & Professionals. Available at: http://www.docstoc.com/search/geoslavery-in-the-stacy-peterson-case/ [accessed 13 Apr 2010]

Dodge, M. and Perkins, C., 2008. Reclaiming the map: British geography and ambivalent cartographic practice, *Environment and Planning A*, 40(6):1271–1276.

Doytsher, Y.; Filin, S. and Ezra, E., 2001. Transformation of datasets in a Linear-based Map Conflation Framework. *Surveying and Land Information Systems*, 61(3):159-169.

Dunkars, M., 2003. Matching of datasets. In *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science (ScanGIS)*. Espoo, Finland. 4-6 Jun 2003.

Edina, 2012. http://edina.ac.uk/digimap/description/products/revisionpolicy.shtml [accessed 24 Jan 2012]

Egenhofer, M.J. and Mark, D.M., 1995. Naïve Geography. In Frank, A. U. and Kuhn, W., eds, *Spatial Information Theory: A Theoretical Basis for GIS*. Berlin: Springer-Verlag, Lecture Notes in Computer Sciences No. 988, p. 1-15.

Ehreke, N., 2006. *Explicit and Implicit Metadata*. [Online]. Available at: http://www.oreillynet.com/onjava/blog/2006/05/explicit_and_implicit_metadata_1.html [accessed 8 Apr 2010]

Elwood, S., 2008. Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, 72:173–183.

Elwood, S., 2009. Geographic information science: new geovisualization technologies – emerging questions and linkages with GIScience research. *Progress in Human Geography,* 33(2):256–263.

EPSRC, 2011. http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/I001212/1 [accessed 23 May 2011]

ESA, 2010. http://www.esa.int/esaNA/SEM2HGF280G_egnos_0.html [accessed 10 Mar 2010]

ESRI, 2005. *GIS Topology*. [Online]. Available at http://www.esri.com/library/whitepapers/pdfs/gis_topology.pdf [accessed 25 May 2012]

Europa Information Society, 2010. http://ec.europa.eu/information_society/policy/psi/library/index_en.htm [accessed 22 Apr 2010]

EuroSDR, 2009. http://www.eurosdr.net/workshops/crowdsourcing_2009/eurosdr_crowdsourcing_2009_repoRt.pdf [accessed 23 Apr 2010]

FGDC, 1998a. FGDC-STD-007.1-1998 *Geospatial positioning accuracy standards. Part 1: Reporting methodology*. Virginia, USA: Federal Geographic Data Committee, [Online]. Available at: http://www.fgdc.gov/standards/standards_publications/index_html

FGDC, 1998b. FGDC-STD-007.3-1998 *Geospatial positioning accuracy standards. Part 3: National standard for spatial data accuracy.* Virginia, USA: Federal Geographic Data Committee, [Online]. Available at: http://www.fgdc.gov/standards/standards_publications/index_html

Fisher, P.; Comber, A. and Wadsworth, R., 2006. Approaches to uncertainty in Spatial Data. In R. Devillers and R. Jeansoulin, eds. *Fundamentals of Spatial Data Quality*. London: ISTE Ltd. Ch. 3, p.43-59.

Flanagin, A. J., and Metzger, M. J., 2008. The credibility of volunteered geographic information. *GeoJournal,* 72:137-148.

Flash Earth, 2010. http://www.flashearth.com/ [accessed 6 Apr 2010]

Flickr, 2010a. http://www.Flickr.com/about/ [accessed 7 Apr 2010]

Flickr, 2010b. http://www.Flickr.com/photos/christophera/3366764093/ [accessed 9 Apr 2010]

Gabay, Y.; Doytsher, Y., 2000. An Approach to Matching Lines in Partly Similar Engineering Maps *Geomatica*, 54(3):297-310.

Garmin, 2010. http://www8.garmin.com/aboutGPS/waas.html [accessed 11 Mar 2010]

geofabrik, 2010. http://download.geofabrik.de/ [accessed 20 Oct 2010]

Girres, J-F. and Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS,* 14(4):435-459.

Goodchild, M.F. and Hunter, G.J., 1997. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11(3):299-306.

Goodchild, M.F., 1993. Data models and data quality: problems and prospects. In M.F. Goodchild, B.O. Parks and L.T. Steyaert, eds, *Environmental Modeling with GIS*. New York: Oxford University Press, p. 141–149.

Goodchild, M.F., 2002. Theoretical Models for Uncertain GIS. In W. Shi, P.F. Fisher and M.F. Goodchild, eds. *Spatial Data Quality*. London: Taylor & Francis. Introduction to Part I & ch. 1, p.1-17.

Goodchild, M.F., 2007a. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2:24–32.

Goodchild, M.F., 2007b. Citizens as sensors: the world of volunteered geography. *GeoJournal*, [Online]. 69:211–221.

Goodchild, M.F., 2008a. Assertion and Authority: The Science of User-Generated Geographic Content. Proceedings of the *Colloquium for Andrew U. Frank's 60th Birthday*. GeoInfo 39. Department of Geoinformation and Cartography, Vienna University of Technology.

Goodchild, M.F., 2008b. Commentary: whither VGI? *GeoJournal* [Online]. 72:239–244.

Google Map Maker, 2010. https://services.google.com/fb/forms/mapmakerdatadownload/ [accessed 6 Dec 2010]

Google Map Maker, 2011. http://www.google.com/support/mapmaker/bin/answer.py?answer=155415 [accessed 1 Oct 2011]

Google Map Maker, 2012. http://support.google.com/mapmaker/bin/static.py?hl=en&page=guide.cs&guide=30028&topic=1094318&answer=155415 [accessed 24 May 2012]

Google Maps API, 2012. http://code.google.com/apis/maps/signup.html [Accessed 24 Jan 2012]

Google Maps Mania, 2010. http://googlemapsmania.blogspot.com/ [accessed 24 Feb 2010]

Groves, P., 2009. *Introduction to GNSS.* [Lecture Notes], Mapping Science, UCL

Guélat, J-C., 2009. Integration of user generated content into national databases - Revision workflow at Swisstopo. [Presentation] In EuroSDR 2009: *Crowd sourcing for updating national databases - a first workshop.* Federal Office of Topography Swisstopo, Wabern, Switzerland. 20-21 Aug 2009.

Haklay, M. and Weber, P., 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing,* 7(4):12-18.

Haklay, M., 2007a. OSM and the public – what barriers need to be crossed? [slide presentation and audio recording] In *SOTM (State of the Map)*. Manchester, UK. 15 Jul 2007 [Online]. Available at: http://sotm2007recordings.blogspot.com/2007/07/muki-haklay-osm-and-public-what.html [accessed 15 Mar 2010]

Haklay, M., 2007b. Democratisation in Web 2.0 and the participation inequality. Po Ve Sham – *Muki Haklay's personal blog*. Available at: http://povesham.wordpress.com/2007/11/14/democratisation-in-web-20-and-the-participation-inequality/ [accessed 9 Apr 2010]

Haklay, M., 2008. Open Knowledge Conference (OKCon) 2008 Presentation. Po Ve Sham – *Muki Haklay's personal blog*. Available at: http://povesham.wordpress.com/2008/03/ [accessed 13 Apr 2010]

Haklay, M., 2010a. *Understanding the quality of user generated mapping –comparing OpenStreetMap to Ordnance Survey geodata*. [Lecture Notes], Mapping Science, University College London.

Haklay, M., 2010b. Usability of VGI in Haiti earthquake response – preliminary thoughts. [Presentation] In *Second Workshop on Usability of Geographic Information.* University College London, UK. 23 Mar 2010.

Haklay, M., 2010c. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England, In *Environment and Planning*, 37(4):682-703.

Haklay, M., Basiouka, S., Antoniou, V. and Ather, A., 2010. How Many Volunteers Does It Take To Map An Area Well? The validity of Linus' law to Volunteered Geographic Information. *The Cartographic Journal* 47(4):315-322.

Harding, J., 2006. Vector Data Quality: A Data Provider's Perspective. In R. Devillers and R. Jeansoulin, eds. *Fundamentals of Spatial Data Quality*. London: ISTE Ltd. Ch. 8, p.141-159.

Harrison, C. and Haklay, M., 2002. The potential of public participation GIS in UK environmental planning: appraisals by active publics, *International Journal for Environmental Planning and Management,* 45 (6) 841-864.

Helft, M., 2009. Online Maps: Everyman Offers New Directions. *The New York Times*, [Online] 17 Nov. Available at: http://www.nytimes.com/2009/11/17/technology/internet/17maps.html?_r=1 [Accessed 19 Apr 2010]

Heo, J.; Kim, J.W.; Park, J.S. and Sohn, H-G., 2008. New Line Accuracy Assessment Methodology using Nonlinear Least-Squares Estimation. *Journal of Surveying Engineering,* 134(1):13-20.

Housingmaps.com, 2012. http://mashupguide.net/1.0/html/ch01s02.xhtml [Accessed 23 May 2012]

Howe, J., 2006. The Rise of Crowdsourcing. Wired magazine June 2006, [Online]. Available at: http://www.wired.com/wired/archive/14.06/crowds.html#Replay [accessed 5 Mar 2010]

Hunter, G. J.; Caetano, M. and Goodchild, M. F., 1995. A methodology for reporting uncertainty in spatial database products. *Journal of the Urban and Regional Information Systems Association*, 7(2):11- 21.

Hunter, G.J., 1999. New Tools for Handling Spatial Data Quality: Moving from Academic Concepts to Practical Reality. *URISA Journal,* 11(2):25-34.

Iliffe, M., 2011. When Government 2.0 Doesn't exist: Mapping Services In The Developing World. In AGI (Association for Geographic Information), 2011. *AGI GeoCommunity Conference 2011 - Placing Ourselves in the New Economy.* Nottingham, U.K. 20-22 Sep 2011.

ISO/TC 211, 2010. http://www.isotc211.org/hmmg/HTML/root.html [accessed 19 Apr 2010]

Jakobsson, A., 2002. Data Quality and Quality Management – Examples of Quality Evaluation Procedures and Quality Management in European National Mapping Agencies. In W. Shi, P.F. Fisher and M.F. Goodchild, eds. *Spatial Data Quality*. London: Taylor & Francis. Ch. 15, p.216-229.

Keogh, A. and Fraser, D., 2008. *Contemporary map products and their origins.* Commission on Education and Training (CET) – ICA 2007-2011. [Online]. Available at: http://lazarus.elte.hu/cet/academic/df-2008-1.pdf [accessed 3 Mar 2010]

Kiiveri, H.T. 1997. Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, 11(1):33-52.

Koukoletsos, T., Haklay, M. And Ellul, C., 2011. An automated method to assess Data Completeness and Positional Accuracy of OpenStreetMap. Presented at *The 11th International Conference on GeoComputation*, London, UK, 20-22 Jul 2011.

Koukoletsos, T.; Haklay, M. and Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS* (in press - DOI: 10.1111/j.1467-9671.2012.01304.x)

Kounadi, O., 2009. *Assessing the quality of OpenStreetMap data.* MSc, University College London, UK.

Kraak, M. J. and Brown, A., 2001. *Web Cartography, developments and prospects*. 1st ed. London: Taylor and Francis Inc.

Kraak, M. J. and Ormeling, F.J., 1996. *Cartography: visualisation of spatial data*. 1st ed. Harlow, Addison Wesley, Longman, 1996.

Leung, Y. and Yan, J., 1998. A locational Error model for spatial features. *International Journal of Geographical Information Science*, [Online]. 12(6):607-620.

Longley, P.A.; Goodchild, M.F.; Maguire, D.J. and Rhind, D.W., 2001. *Geographic Information Systems and Science*. Chichester: Wiley.

Ludwig, I., Voss, A. and Krause-Traudes, M., 2010. How Good is OSM? - Method and Results for Germany. In *Sixth International Conference on Geographic Information Science 2010*, Zurich, Switzerland 14-17 Sep 2010.

Mac Gillavry, 2009. Flirt-ing with the music industry. In AGI (Association for Geographic Information), 2009. *AGI GeoCommunity Conference 2009 - Realising the Value of Place.* Stratford, U.K. 22-24 Sep 2009.

Majchrzak, A., 2011. Emergency! Web 2.0 to the Rescue! *Communications of the ACM*, 54(4).

Mantel D. and Lipeck, U., 2004. Matching Cartographic Objects in Spatial Databases. In *Proceedings of International Society for Photogrammetry and Remote Sensing (ISPRS)*. Vol. XXXV, 35(B4):172–176. Istanbul, Turkey. 12-23 Jul 2004.

Mapping for Change, 2011. http://www.mappingforchange.org.uk/ [accessed 30 Oct 2011]

Maué, P. and Schade, S., 2008. Quality Of Geographic Information Patchworks. In *11th AGILE International Conference on Geographic Information Science 2008*, Girona, Spain. 5-8 May 2008.

McConchie, A.L., 2008. *Mapping Mashups: Participation, Collaboration And Critique On The World Wide Web*. MSc, University of British Columbia.

McDougall, K., 2009. Volunteered Geographic Information for building SDI. In *Surveying and Spatial Sciences Institute Biennial International Conference 2009.* Adelaide, Australia 28 Sep - 2 Oct 2009.

Monmonier, M., 1996. *How to Lie with Maps*. 2nd ed. 1st ed. University of Chicago Press, Chicago, IL.

Mooney, P.; Corcoran, P. and Winstanley, A., 2010. Towards Quality Metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. San Jose, USA. 2-5 Nov 2010.

Mullins, J., 2010. How crowd-sourcing has helped in Haiti. *The New Scientist*, 205 (2745), pp.8-9.

Mummidi, L. and Krumm, J., 2008. Discovering points of interest from users' map annotations. *GeoJournal,* 72:215–227.

Mustière, S. and Devogele, T., 2008. Matching Networks with Different Levels of Detail. *Geoinformatica*, 12(4):435–453.

Nakos, B.; Gaffuri, J. and Mustière, S., 2008. A transition from simplification to generalisation of natural occurring lines. In ICA (International Cartographic Association), *11th ICA Workshop on Generalisation and Multiple Representation*. Montpellier, France 20-21 Jun 2008.

National Map Corps, 2010. http://nationalmap.gov/tnm_corps.html [accessed 22 Apr 2010]

National Map Corps, 2011. http://nationalmap.gov/TheNationalMapCorps/index.html [accessed 28 Oct 2011]

Nielsen, J., 2006. Participation Inequality: Encouraging More Users to Contribute. *Useit.com Alertbox* [Online] Available at: http://www.useit.com/alertbox/participation_inequality.html [accessed 8 Apr 2010]

NLS, 2012. http://www.maanmittauslaitos.fi/en [accessed 9 Feb 2012]

Nov, O., 2007. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64.

O'Brien, O., 2009. Creating and Maintaining Street Orienteering Maps using OpenStreetMap. In GISRUK, *17th Geographical Information Science Research Conference*, Durham, UK 1-3 Apr 2009.

Obermeyer, N., 2007. Thoughts On "Volunteered (Geo)Slavery". [Position Paper] In *Workshop on Volunteered Geographic Information*. University of California, Santa Barbara, USA. 13-14 Dec 2007.

Onsrud, H. and Craglia, M., 2003. Introduction to special issues on access and participatory approaches in using geographic information. *The URISA Journal* 15, APA I, pp. 5-7

Open Source Initiative, 2012. http://opensource.org/ [Accessed 24 Jan 2012]

OpenStreetMap, 2010a. http://wiki.openstreetmap.org/wiki/Statistics [accessed 27 Feb 2010]

OpenStreetMap, 2010b. http://wiki.openstreetmap.org/wiki/Copyright_infringement [accessed 26 Feb 2010]

OpenStreetMap, 2010c. http://wiki.openstreetmap.org/wiki/OpenStreetMap_License [accessed 26 Feb 2010]

OpenStreetMap, 2010d. http://wiki.openstreetmap.org/wiki/Map_Features [accessed 26 Feb 2010]

OpenStreetMap, 2010e. http://wiki.openstreetmap.org/wiki/Main_Page [accessed 1 Mar 2010]

OpenStreetMap, 2010f. http://wiki.openstreetmap.org/wiki/Category:Data_standards [accessed 4 Mar 2010]

OpenStreetMap, 2010g. http://wiki.openstreetmap.org/wiki/Yahoo!_Aerial_Imagery [accessed 4 Mar 2010]

OpenStreetMap, 2010h. http://wiki.openstreetmap.org/wiki/Editor [accessed 8 Mar 2010]

OpenStreetMap, 2010i. http://wiki.openstreetmap.org/wiki/Export#Other_export_formats_and_tools [accessed 8 Mar 2010]

OpenStreetMap, 2010j. http://wiki.openstreetmap.org/wiki/Applications [accessed 9 Mar 2010]

OpenStreetMap, 2010k. http://wiki.openstreetmap.org/wiki/Quality_Assurance [accessed 14 Apr 2010]

OpenStreetMap, 2010o. http://wiki.openstreetmap.org/wiki/Yahoo%21_Aerial_Imagery/Accuracy [accessed 11 Mar 2010]

OpenStreetMap, 2011. Name finder: Abbreviations. WWW document, http://wiki.openstreetmap.org/wiki/Name_finder:Abbreviations

OpenStreetMap, 2012a. http://wiki.openstreetmap.org/wiki/Copyright_Easter_Eggs [accessed 11 Jan 2012]

OpenStreetMap, 2012b. http://wiki.openstreetmap.org/wiki/Open_Database_License [accessed 23 May 2012]

OpenStreetMap, 2012c. http://www.openstreetmap.org/copyright [accessed 23 May 2012]

OPSI, 2010. http://www.opsi.gov.uk/Acts/acts2000/ukpga_20000036_en_1 [accessed 22 Apr 2010]

Ordnance Survey, 2009a. http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/userguides/docs/ITNuserguide.pdf [accessed 15 Jun 2010]

Ordnance Survey, 2009b. http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/userguides/docs/ITNtechspec.pdf [accessed 15 Jun 2010]

Ordnance Survey, 2009c. http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/userguides/docs/OSMMTopoLayerUserGuide.pdf [accessed 15 Jun 2010]

Ordnance Survey, 2010a. http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/ [accessed 13 Apr 2010]

Ordnance Survey, 2010b. http://explore.ordnancesurvey.co.uk/ [accessed 22 Apr 2010]

Ordnance Survey, 2011. http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/research/vernacular.html [accessed 17 May 2011]

Ordnance Survey, 2012. http://www.ordnancesurvey.co.uk/oswebsite/products/os-mastermap/topography-layer/index.html [accessed 17 Jan 2012]

O'Reilly, T., 2005. *What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. *International Journal of Digital Economics* 65:17-37.

OSM wiki pages, 2011. http://wiki.openstreetmap.org/wiki/Isle_of_Wight_workshop_2006 [Accessed 16 Dec 2011]

Ostermann, F; Spinsanti, L, 2011. *A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic***.** *Information For Crisis Management.* In *14th AGILE International Conference on Geographic Information Science 2011*, Utrecht, Netherlands. 18-21 April 2011.

Parker, C.J.; May, A. and Mitchell, V., 2010. An Exploration of Volunteered Geographic Information Stakeholders. In *Proceedings from GIS Research UK 18th Annual Conference.* p. 137-142. University College London, UK. 14-16 Apr 2010.

PC Magazine, 2012. http://www.pcmag.com/encyclopedia_term/0,2542,t=application+programming+interface&i=37856,00.asp [accessed 23 May 2010]

People's Map, 2010. http://peoplesmap.com/About.aspx [accessed 28 Apr 2010]

People's Map, 2012. http://www2.getmapping.com/Products/Peoples-Map [accessed 05 Mar 2012]

People's Place Names, 2010. http://www.yourplacenames.com/ [accessed 22 Apr 2010]

Perkins, C. and Dodge, M., 2008. The Potential of User-Generated Cartography: A Case Study of the OpenStreetMap Project and Manchester Mapping Party. *North West Geography*, 8(1):19–32.

PHP online manual, 2011a. http://php.net/manual/en/function.levenshtein.php [accessed 30 Oct 2011]

PHP online manual, 2011b. http://www.php.net/manual/en/ref.strings.php [accessed 17 Jan 2011]

PhtoSM, 2010. http://haiti.broadbox.de/?zoom=11&lat=18.6036&lon=-72.34315&layers=B00TT [accessed 8 Mar 2010]

Planetdump, 2010. http://wiki.openstreetmap.org/wiki/Planet.osm [accessed 20 Mar 2011]

Poore, B. and Wolf, E., 2010. The Metadata Crisis: Can geographic information be made more usable? [Presentation] In *Second Workshop on Usability of Geographic Information.* University College London, UK. 23 Mar 2010.

programmableweb, 2010. http://www.programmableweb.com/mashups/directory [accessed 25 Feb 2010]

Ramirez, J.R. and Ali, T., 2003. Progress in metrics development to measure positional accuracy of spatial data. In *Proceedings of the 21st International Cartographic Conference (ICC)*. Durban, South Africa. 10-16 Aug 2003.

Ramm, F.; Topf, J. and Chilton, S., 2011. *OpenStreetMap Using and Enhancing the Free Map of the World.* 3rd ed. Cambridge: UIT Cambridge Ltd.

Raymond, E., 1999. The cathedral and the bazaar. *Knowledge, Technology & Policy* 12(3):23-49.

Safra, E. ; Kanza, Y.; Sagiv, Y.; Beeri, C. and Doytsher, Y., 2010. Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. *International Journal of Geographical Information Science*, 24(1):69-106.

Scmitz, S.; Zipf, A. and Neis, P., 2008. *New applications based on collaborative geodata – The Case of Routing*. Publications-Conference Papers, Cartography Research Group, University of Bohn.

Seeger, C.J., 2008. The role of facilitated VGI in the landscape planning and site design process. *GeoJournal,* 72:199-213.

Servigne, S.; Lesage, N. and Libourel, T., 2006. Approaches to uncertainty in Spatial Data. In R. Devillers and R. Jeansoulin, eds. *Fundamentals of Spatial Data Quality*. London: ISTE Ltd. Ch. 10, p.179-210.

Shi, W., 1998. A generic statistical approach for modelling error of geometric features in GIS. *International Journal of Geographical Information Science* 12: 131–43.

Sieber, R., 2007. Geoweb for social change. [Position paper for NGCI Workshop on VGI] Santa Barbara, USA, 13-14 December 2007.

Singer, S., 2009. Openstreetmap with PostgreSQL: The free wiki world map & PostgreSQL. [Presentation] in *PGCon 2009*, Ottawa, Canada, 21-22 May 2009.

Sui, D., 2007. Volunteered Geographic Information: A tetradic analysis using Mcluhan's law of the media [Position paper for the Specialist Meeting on Volunteered Geographic Information] Santa Barbara, USA, 13-14 December 2007.

Sui, D., 2008. The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, 32 (1), pp.1-5.

Swisstopo, 2010. http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/freeproducts.html [accessed 22 Apr 2010]

Switch Maps, 2010. http://www.mapchannels.com/SwitchMaps.aspx [accessed 8 Mar 2010]

The Telegraph, 12 Feb 2009. Matthew Moore: Second council bans apostrophes in street signs http://www.telegraph.co.uk/news/newstopics/howaboutthat/4602491/Second-council-bans-apostrophes-in-street-signs.html [accessed 26 May 2011].

The Turner Ink Blog, 30 Jan 2009. Apostrophes in Birmingham street names: shall we deaf it? http://www.turnerink.co.uk/copywriting-blog/apostrophes-birmingham-street-names-deaf/ [accessed 26 May 2011].

TomTom Map Share, 2012. http://www.tomtom.com/en_gb/maps/map-share/ [accessed 30 Jan 2012]

TomTom, 2010. http://www.tomtom.com/page/openLR [accessed 28 Apr 2010]

Tripadvisor, 2010. http://www.tripadvisor.com [accessed 24 Feb 2010]

Tulloch, D.L., 2008. Is VGI participation? From vernal pools to video games. *GeoJournal,* 72: 161–71.

Turner, A.J., 2006. *Introduction to neogeography*. O'Reilly Media Inc., s.l.

Tveite, H. and Langaas, S., 1995. Accuracy Assessments of Geographical Line Data Sets, the Case of the Digital Chart of the World. In Proceedings from *The 5th Scandinavian Research Conference on Geographical Information Systems*. Trondheim, Norway 12-14 Jun 1995.

Tveite, H. and Langaas, S., 1999. An accuracy assessment method for geographical line data sets based on buffering. *International Journal of Geographical Information Science*, 13(1):27-47.

U.N., 2011. http://www.un.org/en/peacekeeping/missions/minustah/ [accessed 1 Oct 2011]

Ueberschlag, A., 2010. *A first assessment of the OpenStreetMap quality in Switzerland*, Unpublished manuscript EPFL, Switzerland.

UKMap, 2011. http://www.geoinformationgroup.co.uk/products/mapping [accessed 30 Oct 2011]

Van Exel, M.; Dias, E. and Fruijtier, S., 2010. The impact of crowdsourcing on spatial data quality indicators. In *Sixth international conference on Geographic Information Science*. Zurich, Switzerland. 14-18 Sep 2010.

Van Niel, T.G. and McVicar, T.R., 2002. Experimental evaluation of positional accuracy estimates from a linear network using point- and line-based testing methods. *International Journal of Geographical Information Science*, 16(5):455-473.

Van Oort, P.V., 2006. *Spatial data quality: from description to application.* PhD, Netherlands Geodetic Commission, Delft, The Netherlands.

Vauglin, F., 2002. A Practical Study on Precision and Resolution in Vector Geographical Databases. In W. Shi, P.F. Fisher and M.F. Goodchild, eds. *Spatial Data Quality*. London: Taylor & Francis. Ch. 9, p.127-139.

Veregin, H., 2000. An accuracy assessment method for geographical line data sets based on buffering. *International Journal of Geographical Information Science*, 14(2):113-130.

Veregin, H., 2005. Data quality parameters. In: Longley, P.A.; Goodchild, M.F.; Maguire, D.J. and Rhind, D.W., eds. *Geographic Information Systems and Science*. 2nd ed. New York: Wiley. Ch. 12, p.177-189

Walking-papers, 2010. http://walking-papers.org/about.php [accessed 8 Apr 2010]

Walsh, J., 2008. The beginning and end of neogeography. *GEOconnexion International Magazine*, 7(4):28-30

Walter, V. and Fritsch, D., 2001. Matching spatial datasets: a statistical approach. *International Journal of Geographical Information Science*, 13(5):445-473.

Wikimapia, 2010a. http://wikimapia.org/wiki/Main_Page [accessed 28 Apr 2010]

Wikimapia, 2010b. http://wikimapia.org/terms_reference.html [accessed 28 Apr 2010]

Wikipedia, 2010. http://www.wikipedia.org [accessed 24 Feb 2010]

Willis, N., 2009. 'Why OpenAerialMap failed where OpenStreetMap succeeded' *LWN.net*. [online] 24 Jun. Available at: http://lwn.net/Articles/338491/ [accessed 19 Apr 2010]

Yourtomtom, 2010. http://www.yourtomtom.com/contact/3/advertise.html [accessed 28 Apr 2010]

Zandbergen, P.A., 2008. Positional Accuracy of Spatial Data-Non-normal distributions and a critique of the National Standard for Spatial Data Accuracy. *Transactions in GIS*, 12(1):103-130.

Zielstra, D. and Zipf, A., 2010. A Comparative Study of Proprietary Geodata and VGI for Germany. In *13th AGILE International Conference on Geographic Information Science*. Guimaraes, Portugal. 10-14 May 2010.

APPENDICES

APPENDIX A: Description of the developed application

A description of the application that supports the method described in this thesis follows through screenshots and a brief explanation of what the developed code does. The code is implemented in PHP in several pages.

Page 1: The first page (filename: '1.php') collects the necessary information so that the user can connect to the spatial database (Figure 1).

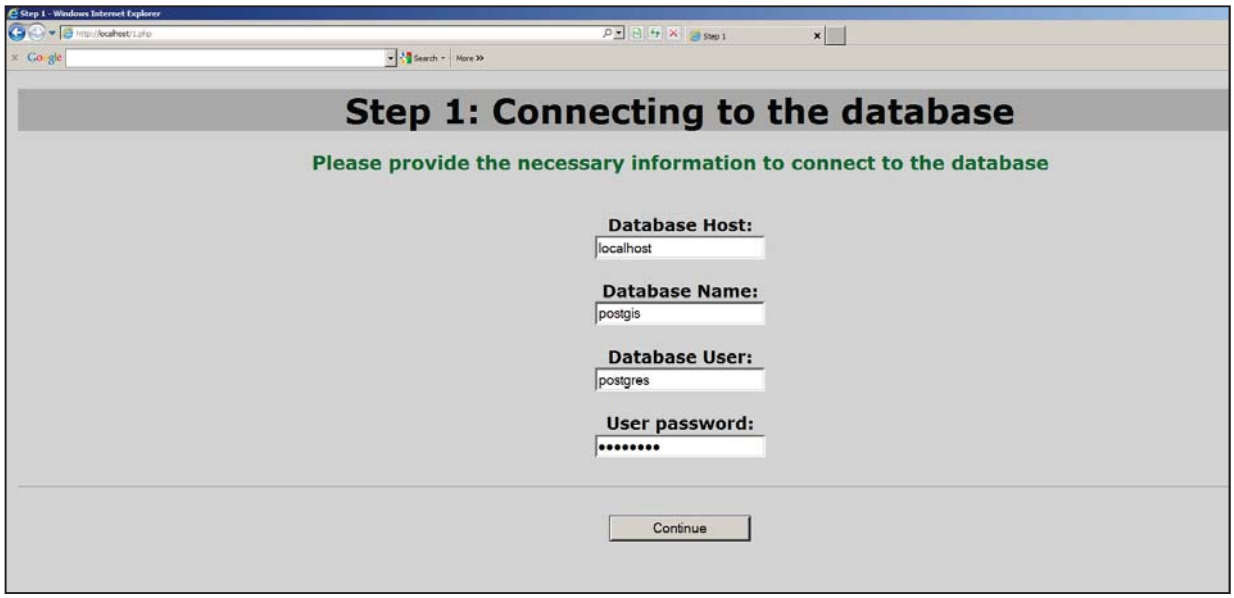

Figure A-1: *First page: Providing the user credentials*

Page 2: If the user credentials are not correct, a relevant message notifies about the database connection failure and the user is directed back to the first page. If the connection is successful, the user can select the datasets and tessellation file from a drop-down list (Figure 2). An additional check is performed for existing tables that will be later produced. If they exist, this means that a previous evaluation is interrupted, finished or currently running, and the database has not been cleared from the intermediate database tables. This is essential for the user to know, because all these existing tables will be deleted and recreated for the new evaluation, so a possible data loss must be prevented.

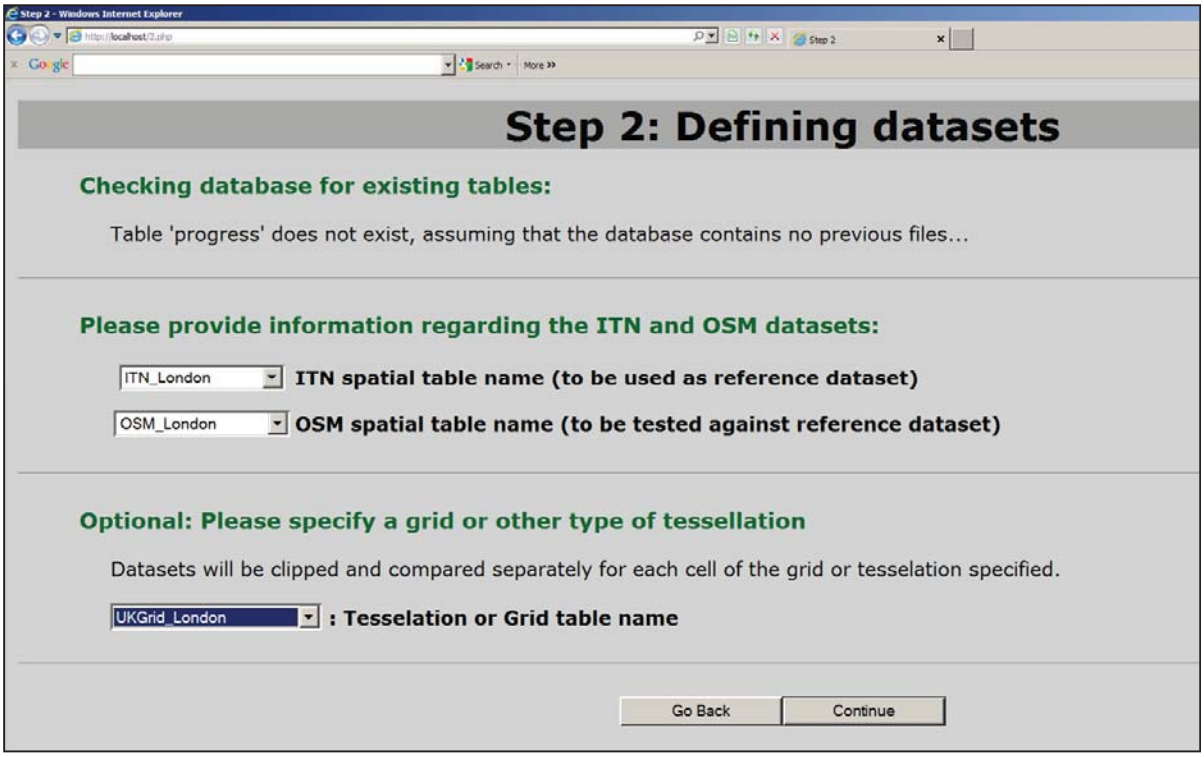

Figure A-2: *Defining datasets*

Page 3: The next page (Figure 3) checks where the geometry is stored for each input dataset and asks for the desired tessellation ID that will be used to identify the tiles. The drop-down list limits the choices to the columns with no duplicate values, which can further be used as a primary key.

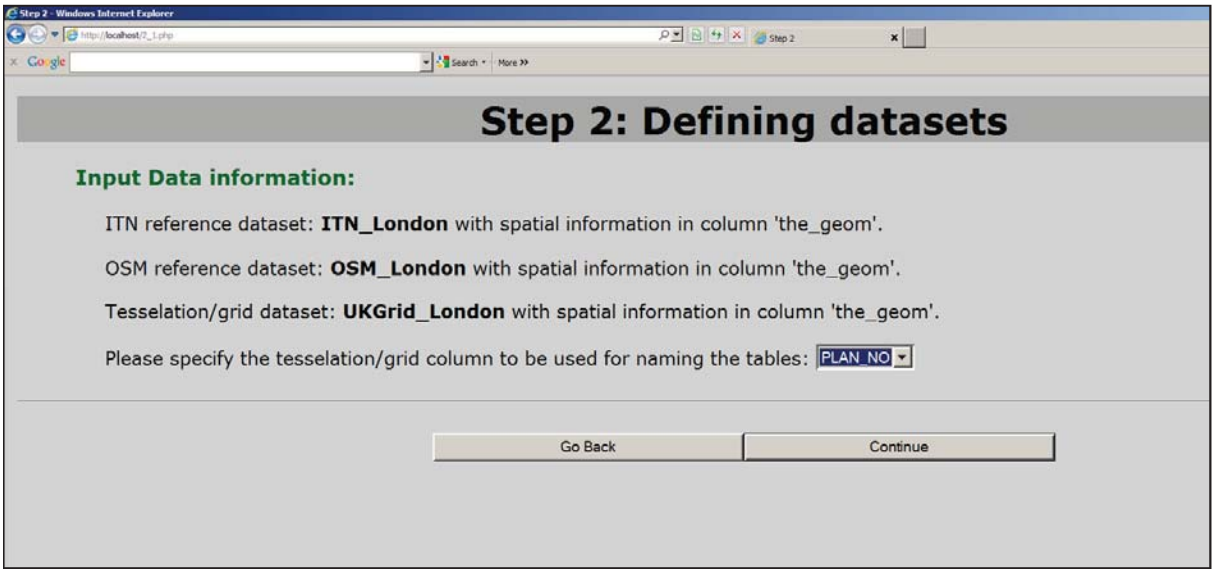

Figure A-3: *Defining tile names*

Page 4: The next page (Figure 4) checks that the coordinate system is common for all three input files (two datasets and tessellation file), as well as if they overlap. In case of different reference systems or no overlapping, the user is notified and directed back to page 2. In case of a partial overlap, the user is notified that the evaluation will be performed only for the commonly described area. A bounding box is created for each input file and a fourth bounding box (the intersection of the three ones) is created to clip the data accordingly, in case of partially overlapping datasets.

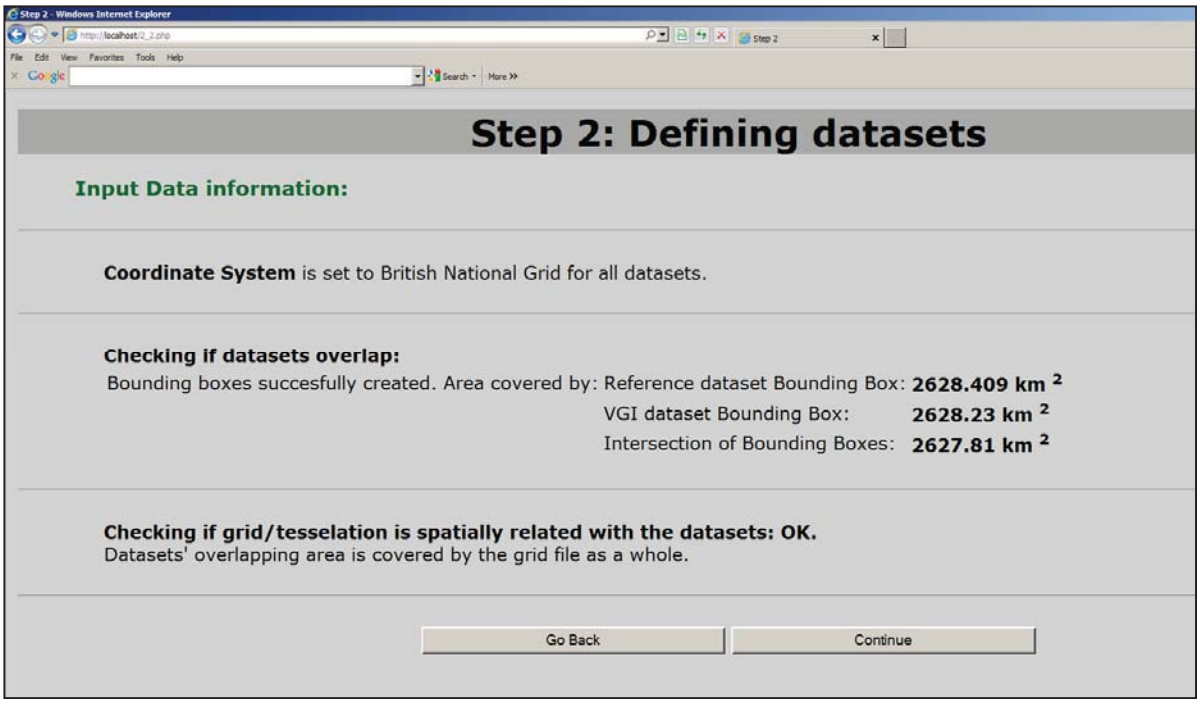

Figure A-4: *Checking datasets consistency*

Page 5: The next page (Figure 5) enables the selective data evaluation, according to the road types that are found in the datasets. By unchecking them, they will be removed from the datasets and will not be examined. A drop-down list is provided for VGI datasets (on the right), as lack of standards can lead to a much richer network classification. Two 'Details' buttons can provide additional information in a pop-up window to aid road type selection or rejection (Figure 6). For example, VGI road types that include only 1 or few features and a limited length (such as 'fence', 'footway;service', 'proposed', 'crossing' of Figure 6 on the right) could be rejected. An additional option is provided if separate output files are needed per tile, however this is not suggested, since a high number of files will be created. In any case, information per tile can also be extracted with a little effort from the final and total results.

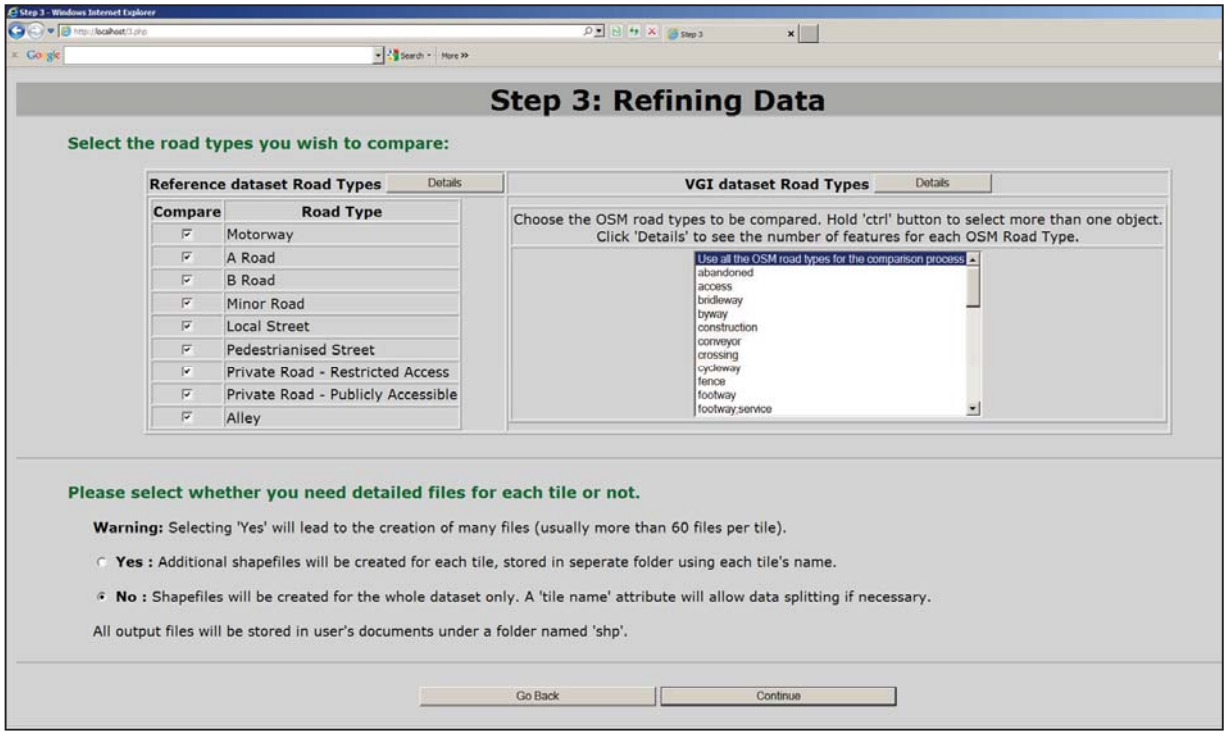

Figure A-5: *Selecting road types*

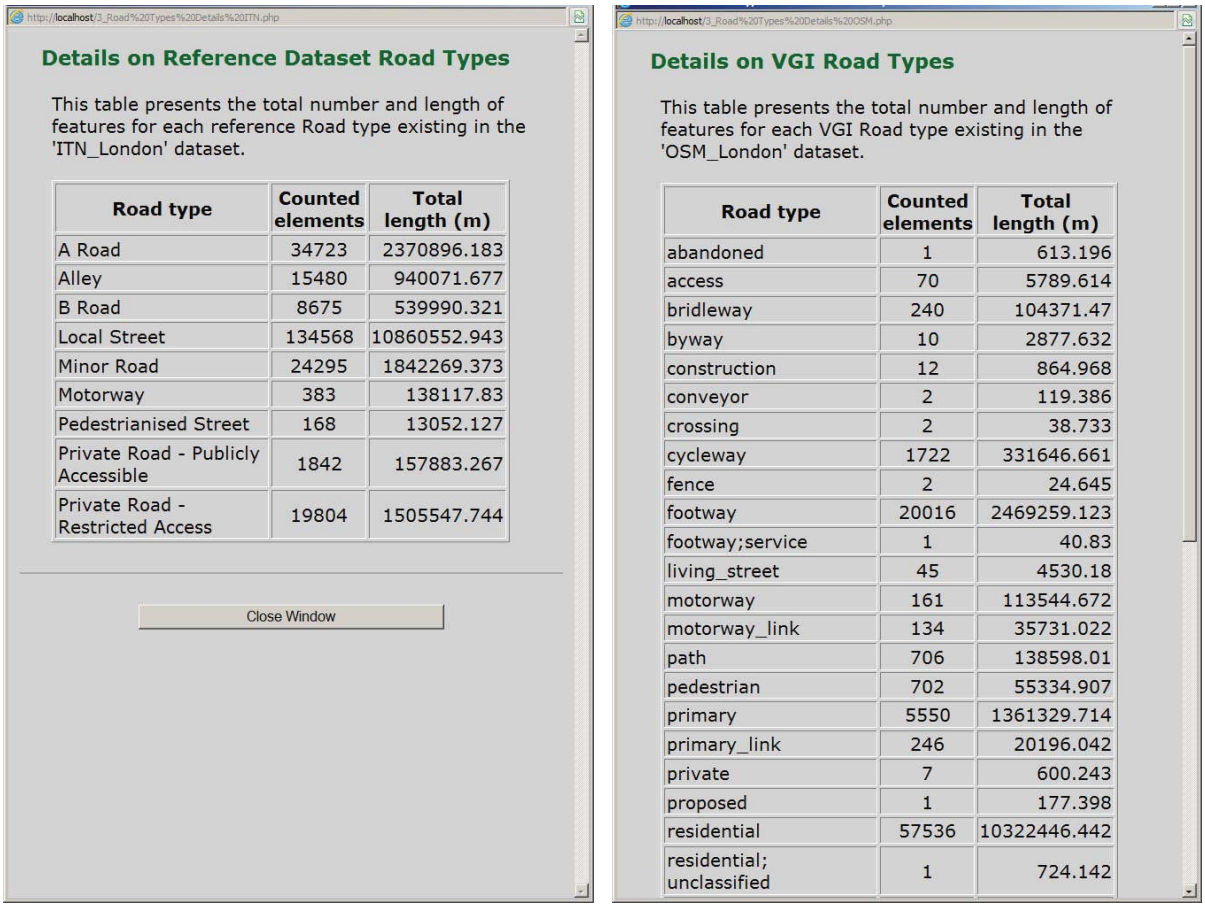

Figure A-6: *Details on reference and VGI road types*

Page 6: New datasets are created, based on the selections of the previous page and the necessary clipping of page 4. Depending on the network density and data volume, this page may take a while to load. As Figure 7 shows, the user needs to define the parameters for the positional accuracy method. There are two options; the simplified IBM version, where the user defines the iterations and step of buffering, and the complex one, where the user defines a level of confidence (overlap percentage). Due to the limitations of the simplified version (discussed in section 4.12), it is not fully supported and the default option is set to the complex one. Both versions require a starting buffer. The last web-page option, regarding PostgreSQL path, deals with differences between 32-bit and 64-bit Operating Systems and needs to be correctly defined, otherwise the output spatial tables will not be automatically exported to shapefiles (however, they could be exported manually afterwards). By pressing 'Continue', the comparison process commences.

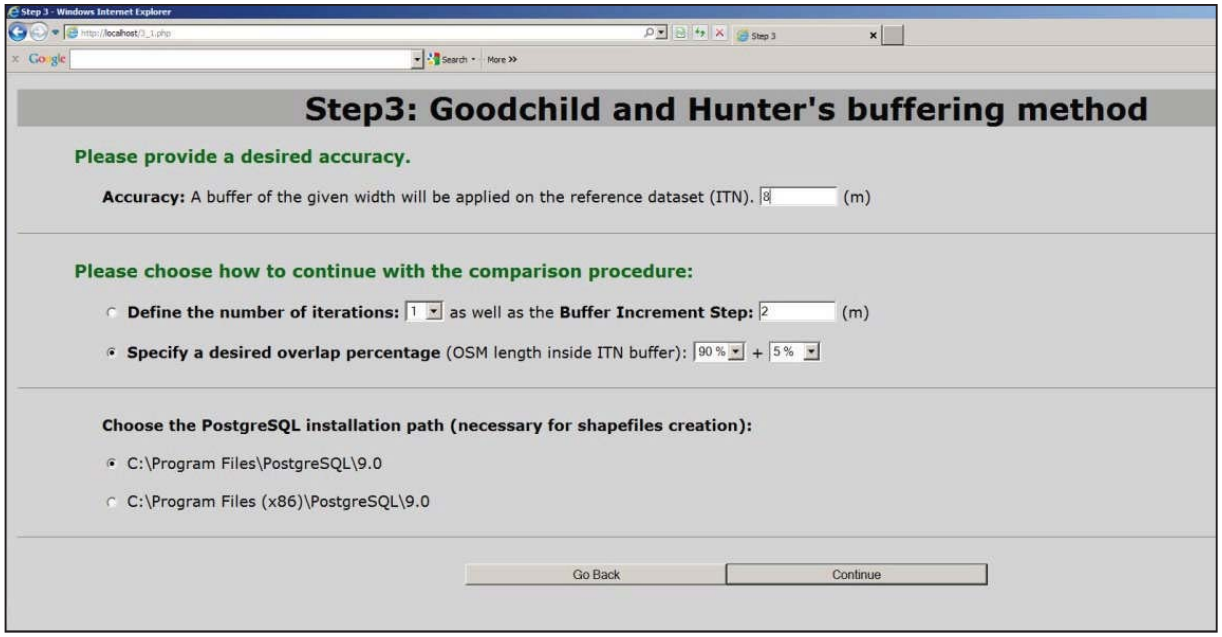

Figure A-7: *Customisation of the positional accuracy approach*

Page 7: While the process is running, the user is notified by a frequently refreshing page about the progress. Each tile is processed individually and data matching, data completeness, attribute and positional accuracy measurements are performed. Figure 8 presents three types of notification, according to the action currently performed.

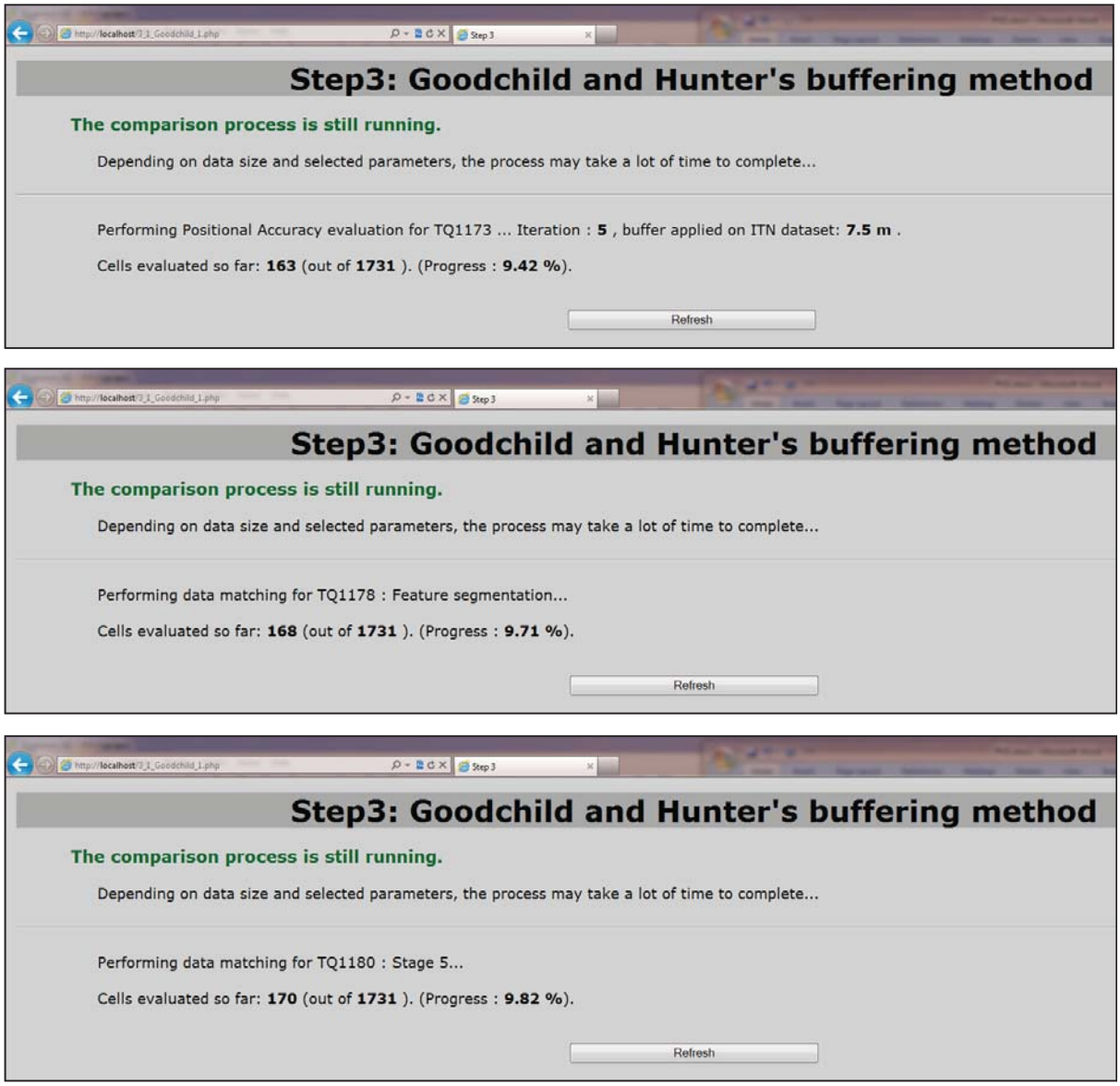

Figure A-8: *Notifications of the comparison progress*

When all tiles are processed, data are aggregated and output files (CSV and shapefiles) are created and exported. When the whole process is finished, a message appears similarly to the ones of Figure 8 and the output files can then be accessed. A brief description of the produced output files follows. The output filenames refer to the ITN and OSM datasets used as an example in Figure 2, however in a general case ITN links to the reference and OSM to the VGI dataset.

Spatial Tables exported as Shapefiles:

- **cells_done:** A polygon spatial table that includes all the tiles that were processed (only geometry information – no quality results, similar structure to the tessellation file used).

- **ITN_match_all**: A linear spatial table that includes the matched features of the reference dataset.
- **OSM_match_all**: Same as above for the VGI dataset.
- **ITN_no_match_all**: A linear spatial table that includes the non-matched features of the reference dataset.
- **OSM_no_match_all**: Same as above for the VGI dataset.
- **ITN_buffer_all**: A polygon spatial table that includes the selected buffer zone (different for each tile) describing its positional accuracy.
- **OSM_Intersected_all**: A linear spatial table that includes the part of the matched VGI dataset which lies inside the above buffer.
- **final_results_m**: A polygon spatial table, created by joining each tile from the cells_done shapefile with the measured quality elements' values. This table contains length information, expressed in meters.
- **final_results_pct**: Same as above but expressed in percentage.

The last two shapefiles contain the full quality information for each tile of the tested area. However, due to an existing shapefile format bug, null values for tiles with no information are considered and represented as zero values, which confuses non-examined tiles with those that are examined but found with zero quality value. Additionally, for large areas (e.g. second case study) the shapefile grows in size and is difficult to be handled by any GIS software. To deal with this problem, additional shapefiles are automatically generated from the 'final_results_pct' shapefile for each quality element individually, containing only the tiles with non-null values in the appropriate column. These are:

- **ITN_data_match**: Reference dataset's matched percentages (showing VGI completeness).
- **OSM_data_match**: VGI dataset's matched percentages (indicating VGI over-completeness).
- **ITN_att1_match:** Reference dataset's attribute percentages for primary name (showing VGI primary name attribute accuracy).
- **ITN_att2_match:** Reference dataset's attribute percentages for secondary name (showing VGI secondary name attribute accuracy).
- **ITN_att_match:** Reference dataset's attribute percentages for both names (showing VGI total attribute accuracy).
- **OSM_att1_match:** VGI dataset's attribute percentages for primary name (indicating VGI primary name over-completeness).
- **OSM_att2_match:** VGI dataset's attribute percentages for secondary name (indicating VGI secondary name over-completeness).

- **OSM_att_match:** VGI dataset's attribute percentages for both names (indicating VGI total attribute over-completeness).
- **OSM_pos_acc:** VGI dataset's positional accuracy.

Additional Spatial Tables of minor importance, also exported as Shapefiles:

- **ITN_segments_all**: A linear spatial table that includes all the segments of the reference dataset. Its purpose is to check and ensure that all features are appropriately segmented.
- **OSM_segments_all**: Same as above for the VGI dataset.
- **ITN_seg_match_all**: A linear spatial table that includes the matched segments of the reference dataset. This refers to data matching stages 1 to 4. Its purpose was to develop and evaluate the segment-by-segment data matching procedure for each stage.
- **OSM_seg_match_all**: Same as above for the VGI dataset.
- **ITN_seg_no_match_all**: A linear spatial table that includes the non-matched segments of the reference dataset. Its purpose was similar to the 'ITN_seg_match_all' shapefile.
- **OSM_seg_no_match_all**: Same as above for the VGI dataset.

Non-spatial Tables exported as CSV files:

- **Matching_Stats_m**: Information for each tile of both datasets regarding the network lengths used during each stage of the matching procedure, as well as for the calculation of data completeness and attribute accuracy (Table 4.5 provides an example).
- **Matching_Stats_pct:** Same as above but expressed in percentage.
- **Matching_pct_ITN:** Same as above (percentages) but only for the reference dataset.
- **Matching_pct_OSM:** Same as above but only for the VGI dataset.
- **Match_all:** Aggregated values of tables 'Matching_Stats_m', Matching_Stats_pct' for the whole area and both datasets, expressed in meters and percentage respectively.
- **Goodchild_Stats:** Information about each tile regarding VGI's positional accuracy (Table 4.6 provides an example).
- **cellresults:** Detailed information about the positional accuracy algorithm performance on each tile, referring to the buffer width and overlap percentage of each iteration.
- **Road_Types_Matching:** Existing pairs of matched road types and their aggregated length.
- **Road_Types_Match_ITN:** Information from the previous table regarding the reference dataset, grouped by reference road type, expressed in percentages and presented in a descending order.
- **Road_Types_Match_OSM:** Same as above for the VGI dataset.

- **Road_Types_notmatched:** Information on the length of road types from both datasets that were not matched at all (if any).
- **Roadname_acc_nm:** Information on features of both datasets, matched and non-matched, with unique primary or secondary road name attributes.
- **WhatMappedITN:** Information about the matched length and percentage for each road type of the reference dataset.
- **WhatMappedOSM:** Same as above for the VGI dataset.

During the evaluation process, the user can have a more comprehensive view of how the comparison and evaluation is performed. QGIS is an open-source application that connects to the postGIS database quite easily. Spatial tables that are currently being processed can be loaded without disturbing the process that runs in the background. By refreshing the data window (which is also done by a simple pan or zoom), the user can visualise how the proposed framework works. Figure 9 provides an example for the datasets selected in Figure 2. Matched and non-matched spatial tables are loaded and the user has access to the data matching results tile-by-tile, simply by refreshing the page. Dark green and red colour are used for reference and VGI matched data respectively. Light green and pink are used for reference and VGI non-matched data respectively. Figure 9b is created by slightly panning the data window of Figure 9a after a few seconds.

Other spatial tables can also be loaded, for example the buffer spatial table will provide an insight of the binary search algorithm by showing how the buffer reshapes and converges to the final value that expresses the positional accuracy. Depending on the data density of the tile and the hardware capabilities, however, this procedure may be completed faster than the user's interaction with the database tables.

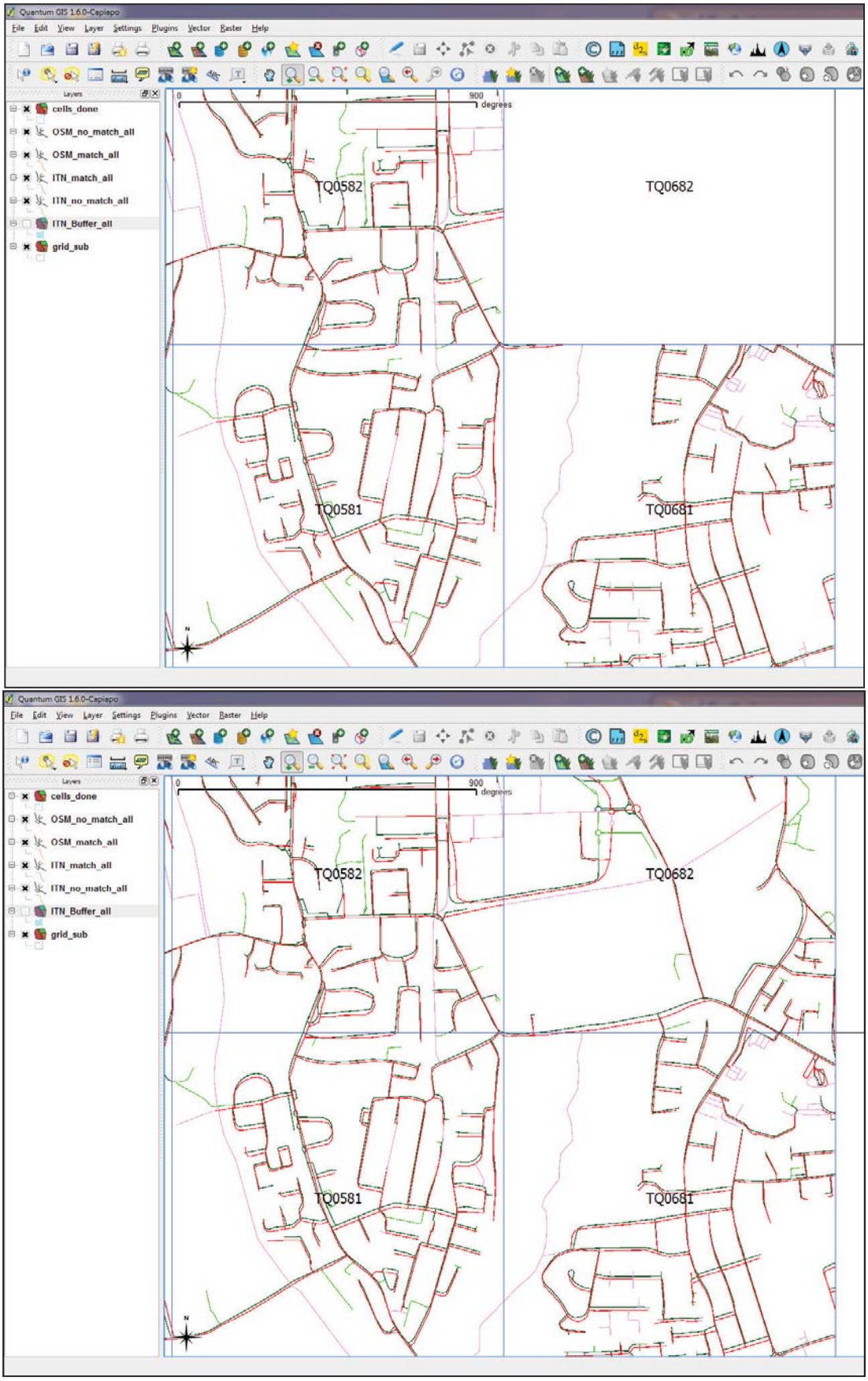

Figure A-9: *Visualisation of the progress using QGIS, a: currently processing TQ0682, b: data refresh after 5 seconds*

APPENDIX B: Other characteristics of VGI

For a broader view of VGI, other characteristics that are not directly related to data quality will be briefly mentioned.

B1. Motivation of VGI

The credibility issue can be partially attributed to the fact that we do not know what motivates VGI participants to contribute. Research on the possible motives already exists, starting from web volunteered contribution in general (e.g. Nov, 2007 for wikipedia), applying the general research findings to VGI (e.g. Coleman *et al.*, 2009) or specifically addressing VGI motivation (Budhathoki *et al.*, 2009). Especially the latter proposes a large list of motivation factors which aims – among others – to provide an insight about the content and quality of VGI.

B2. The digital divide, ethics and values of VGI

The digital divide refers to who creates VGI, who can access it and who is mapped. Although digital divide can also be present in government data resources (Elwood, 2008), the 'volunteered' part of VGI enhances the problem by having distinctive spatial data coverage and usage.

Bruns (2008), although not dealing specifically with GI, defines four categories according to the way the contribution of 'produsers' is exploited: '*Harnessing the Hive*', '*Harvesting the Hive*', *'Harbouring the Hive*', '*Hijacking the Hive'.* Specifically for VGI and based on the purpose of contribution, Haklay (2010a) distinguishes five types of VGI, mentioning corresponding examples, as: '*Egalitarian'*, *'Covert profiteering'*, '*Conspicuous profiteering*', *'Disingenuous'*, and '*Exploitative'.*

VGI can also be used by commercial providers in search for market patterns by collecting information about users' preferences, without them knowing it (Brown, 2001). VGI projects may require personal information to be made public. Dobson and Fisher (2003) define it as 'geoslavery' and warn that the need for safety and security forces a trade-off between provision of spatial data and security, enabling the monitoring and surveillance of people without question and by ignoring potential hazards; Obermeyer (2007) refers to it as 'Volunteered (Geo)Slavery'. Goodchild (2008b) mentions the misuse of VGI to compromise National Security or to harm people who are insufficiently mobile. Dobson (2008) argues on the severity of geoslavery by referring to maybe its first known victim. Sui (2007) argues that protocols and standards need to be established so as to prevent misuse of VGI, especially in the areas of public health and homeland security.

B3. Copyright issues of VGI

The exploitation of contributed data raises the question of copyright issues. While producers with the traditional meaning legally hold copyright in their work, this is not feasible for content 'produsers' (Bruns, 2008). Copyright legislation in the context of a protecting frame is difficult to be part of VGI, mainly because it is collaborative and usually anonymous. Additionally, it is not static; through an iterative procedure the information is continuously altered by different users, so it is difficult to define the person to whom the law applies to. However, a relatively loose copyright context exists in many projects. New copyright schemes are recently developed to cover open source data, under Open Source Initiative (2012), where someone can find a plethora of licenses, review them and chose which one fits a specific purpose. Usually they permit the free use and distribution of data with some restrictions, which often aim to prevent VGI from being used as a closed source or from being claimed to be someone's property (McConchie, 2008), or even used for illegal, immoral, unethical purposes and other activities that do not respect privacy (Google Maps API, 2012). In any case, however, finding the most appropriate license scheme is not an easy task, since a wrong choice may lead to a very strict frame that will limit the use of data, or to a loose one that could allow legal actions for copyright protection and endanger the whole VGI project. On the other hand, as VGI is dynamic and adapts to people's needs, a license scheme may need to be changed, as in the case of OSM (Chilton, 2009b).

B4. Sustainability of VGI

Viability of VGI is a matter of discussion. Although enthusiasts keep rising, no one knows if this will simply be a trend and after a while some or all VGI projects will be deserted, 'steadily growing out-of-date' (Goodchild 2008b; Sui, 2008). A lot of things can drive users away, in many cases unpredictable. There are already examples of VGI projects that failed and no longer exist, such as the raster orientated 'OpenAerialMap' (Willis, 2009). Others are unlikely to succeed, like 'Vernal Pools' (Tulloch, 2008). The development of commercial activity around VGI projects may either drive users away or help the long-term viability of the VGI project (Bruns, 2008). However, as Haklay (2008) argues, the egalitarian model of VGI can be quite complex compared to a techno-libertarian one; while the egalitarian approach links personal benefit with a social payback, the techno-libertarian considers the benefit of one side against another.

APPENDIX C: VGI Commission and other data matching examples

This Appendix presents some examples of VGI commission (over-completeness), as well as other findings of the data matching procedure from all the case studies. Figures 1 and 2 refer to the first case study, Figures 3 to 8 to the second and Figures 9 to 14 to the third and final one. They complement the indicative examples used in each case study (sections 5.5.3, 6.5.3 – see Table 6.24 for more details – and 7.5.3). All figures represent the reference dataset in yellow (regardless of being matched or not) and the VGI non-matched dataset in red (Matched VGI is not visible).

Figure C-1: *London region, OSM commission (parking area at Heathrow airport)*

Figure C-2: *London region, OSM commission*

Figure C-3: *Lancashire region, OSM commission*

Figure C-4: *Severn region: OSM seems to have inconsistent data*

Figure C-5: *Humberside region, OSM commission*

Figure C-6: *Essex region, OSM commission*

Figure C-7: *Yorkshire region, OSM commission*

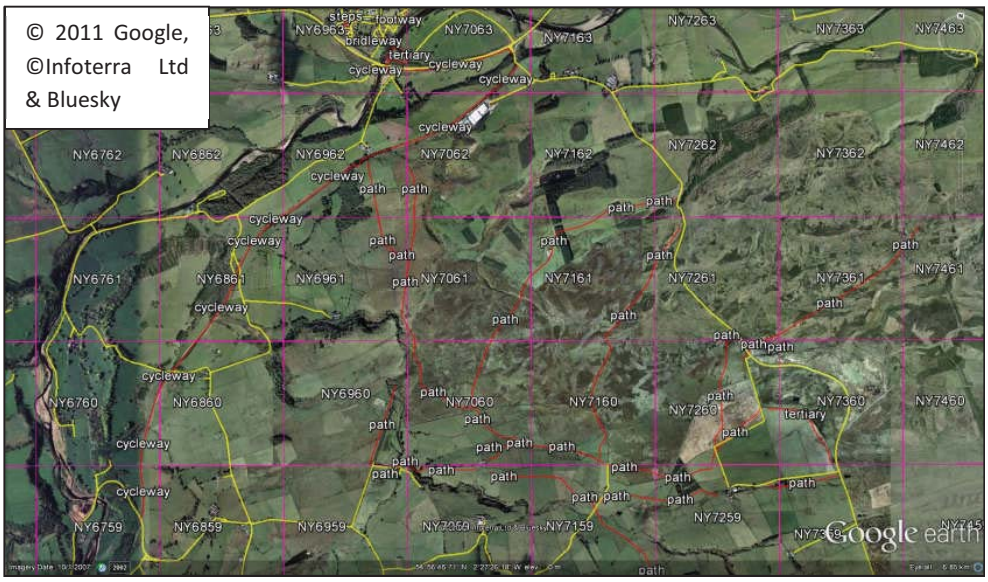

Figure C-8: *North region, routes for pedestrians – bicycles*

Figure C-9: *Haiti area, UN-GMM datasets: GMM commission, example 1*

Figure C-10: *Haiti area, UN-GMM datasets: GMM commission, example 2*

Figure C-11: *Haiti area, UN-GMM datasets: Failed data matching due to distance. Mislocated reference dataset?*

Figure C-12: *Haiti area, UN-GMM datasets: GMM commission, example 3: Non-traffic road type*

Figure C-13: *Haiti area, GMM-OSM datasets: OSM commission, example 1*

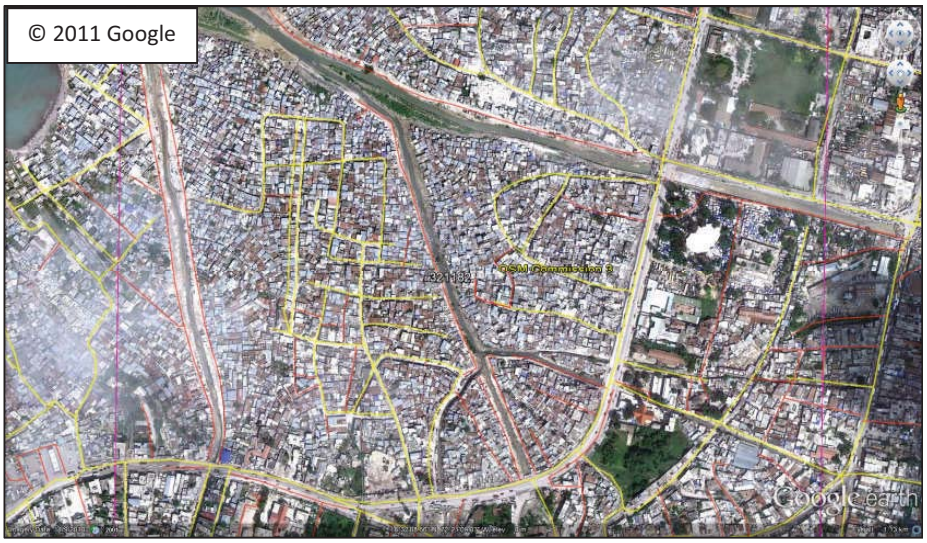

Figure C-14: *Haiti area, GMM-OSM datasets: OSM commission, example 2*

Epilogue

It is difficult to write an epilogue in a long-lasting research that started back in 2009. When contemplating each of its four stages, there are few words to describe the feelings. The first stage demanded a strenuous search for information to develop a general knowledge of VGI and its implications. Gradually the general aspects that needed further research started to form and research questions faintly appeared. This led to a more specified search for publications and other relevant information, roughly shaping the research direction. The decision to move to the next stage was really hard: there was always one more publication, thesis, article or presentation that needed to be studied, which subconsciously might also have been an excuse to avoid moving to the next and unknown stage of the method development.

The method development combined the theoretical conception of an algorithm and the tools to put it into effect. Lack of programming skills demanded some extra effort, which however interacted with the algorithm: the algorithm guided programming development and the trial and error results, from the preliminary to the advanced ones, improved the algorithm. Starting from scratch, one may become obsessed when realizing that slowly but gradually the theoretical model becomes a practical application. Fresh publications or other findings that demanded further study occasionally reminded of the first and theoretical stage, which by then looked far distant and boring. Every possible case scenario had to be predicted and properly handled by the code. The method was developing through trials by altering the code and its parameters and checking for improvements and deteriorations. There was always something that somehow needed to be improved, which, similarly towards the end of the previous stage, prevented moving on to the next stage.

The third stage refers to the application of the method on several case studies, collection of the results, evaluation and interpretation. This stage was more directly linked to the previous one, as unpredicted cases required some corrections in the code and repetition of the analysis. The benefits of developing an automated method started to appear every time a repetition was necessary. Soon errors subsided to anticipated or easy-to-justify levels, allowing for further interpretation of the results. The satisfaction of finding the method efficient and robust leads to its next application or case study. New and interesting findings in each case increase the eagerness to apply it elsewhere, postponing the final stage, similarly as towards the end of the previous stages.

The final stage of writing up comes when someone realizes (or is forced to realize) that research can be endless, since each answer creates a new question. All the gained knowledge had to be placed in order, keeping the parts that form a logical series of arguments for the final thesis. There was a lot of information that had to be rejected: knowledge from publications or other sources during the first stage not exactly related to the research questions and objectives, failed attempts during the method development of the second stage, results and findings from the method application considered not relevant or important. On the other hand, information from recent publications had to be filtered and added to enrich and update the thesis. Putting it together was not an easy task.

Regardless of someone's determination and dedication to finish a PhD, all these are not possible without the appropriate supervision. A successful supervisor does not only provide guidance within each stage, but knows when to shake, push or force the student to move to the next stage. In the writer's opinion, getting lost in one of the above stages is very easy.

Entering the research area back in 2009 as a newbie, one thought is gradually formed about research, also applying to measurements in my familiar land surveying area: despite hard efforts, nothing can be perfect in one research. What seems perfect, however, is to know about the things that are not perfect, why they are not perfect and what could potentially correct or partially improve them. The continuous effort for improvement, either singularly or collaboratively, either by the same researcher or by another one in the future, is the reason why research never ends.